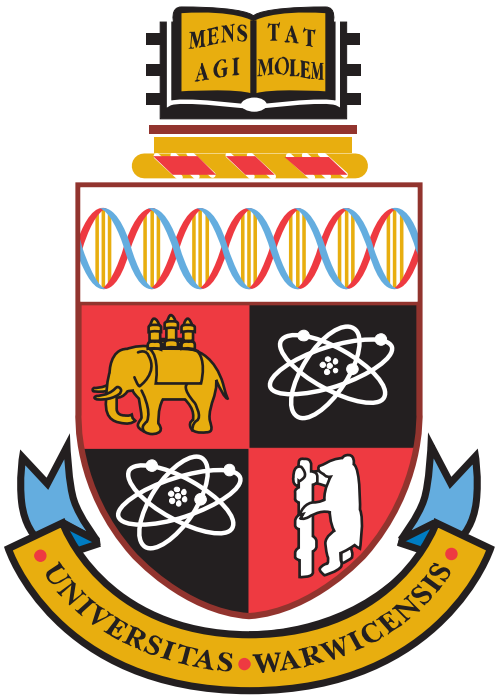
The demarcation of transcription factor binding sites through the analysis of DNase-seq data

Jason Piper

Systems Biology Doctoral Training Centre
University of Warwick



A thesis submitted for the degree of Doctor of Philosophy
in Systems Biology

Supervisors: Sascha Ott and Constanze Bonifer

December 2014

Contents

Acknowledgments **i**

Declaration **iii**

Abstract **v**

Abbreviations and Acronyms **vii**

1 Introduction **1**

1.1 Motivation 1

1.2 Overview 2

1.3 Background 3

1.3.1 Transcriptional regulation 3

1.3.2 The chromatin landscape 3

1.3.3 Regulation from a distance 4

1.3.4 Characterising regulatory elements 5

1.4 Identifying transcription factor binding sites 9

1.4.1 Chromatin immunoprecipitation 9

1.4.2 DNase I footprinting 12

1.4.3 DNase-seq for identifying transcription factor binding sites . 14

1.5 Analysing transcription factor binding data 18

1.5.1 Peak calling 18

1.5.2 Digital genomic footprinting 19

1.5.3 Identifying DNA binding motifs 20

1.6 Introduction to the thesis 22

2 Footprinting analysis of DNase-seq data **25**

2.1 Motivation 25

2.2 Contributions 26

2.3 Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data . . . 26
2.4 Supplemental material . . . 39
2.5 pyDNase documentation . . . 76
2.6 Corrigendum . . . 89

3 Differential DNase-seq footprinting 91
3.1 Motivation . . . 91
3.2 Contributions . . . 92
3.3 Differential DNase-seq footprinting identifies cell-type determining transcription factors . . . 92
3.4 Supplemental material . . . 101
3.5 pyDNase 0.2.0 footprinting tutorial . . . 108

4 Discussion 119
4.1 Footprinting analysis of DNase-seq data . . . 119
4.2 Application of Wellington to clinical samples . . . 122
4.3 Differential DNase-seq footprinting . . . 123
4.4 Outlook . . . 126
4.5 Conclusions . . . 129

Bibliography 131

Appendix A Identification of a dynamic core transcriptional network in t(8;21) AML regulating differentiation block and self-renewal 147

Acknowledgments

I would like to thank my supervisors Sascha and Conny, for their support and unfaltering tolerance of my energetic tendencies over the last four years. I would also like to thank (in no specific order) Hugo van den Berg, Peter Cockerill, Pierre Cauchy, Nigel Dyer, Peter Krusche, Markus Elze and Salam Assi, who were also critical to my academic development and the overall success of the project. And to Anne Maynard, Sarah Shute, and Brent Kiernan, who kept the cogs turning and the funding flowing.

Words cannot explain how much love I have for my fellow PhD students and housemates Jess, Markus, Sam, and Jess, (and Tom, for a bit, at least) who endured late night discussions about ChIP-seq, motifs, DNase-seq, and other things completely irrelevant to their own work. Along with my other new friends - Shona, Claudia, Kat C, Cat S, Kate R, Paul H and everyone in the Systems Biology and MOAC DTCs, they are without a doubt some of the best people I have the pleasure of being friends with. This chapter of my life was far harder to finish than any of the chapters presented here.

To old friends (you know who you are) that have stuck around even when I disappeared for 4 years to do 'something to do with DNA.' I'm very sorry for missing the Avondale Road mince pie party.

And to my family, who never seem to be too far away (for better or for worse), and to whom this thesis is dedicated. Everything here has only been possible because of all you've done for me.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. A 'Contributions' section prefaces each results chapter, outlining the work performed by myself, and contributions from collaborators.

Abstract

The expression of eukaryotic genes is controlled by non-coding regulatory elements such as promoters and enhancers, which bind sequence-specific DNA-binding proteins (transcription factors). In multicellular organisms, the characterisation of these elements is required in order to understand how a single genome is utilised to generate a multitude of cell types, and how aberrant regulation of transcription contributes to disease processes. This involves the identification of transcription factor binding sites within regulatory elements that are occupied in a defined regulatory context. Digestion with DNase I and the subsequent analysis of regions protected from digestion followed by high-throughput sequencing (DNase-seq footprinting), allows for the quantification of genome-wide transcription factor binding. However, the handful of methods for analysing DNase-seq data has not been extensively validated or benchmarked. This thesis describes a novel footprinting algorithm, Wellington, which is presented in the context of a comprehensive comparison of several other DNase-seq footprinting algorithms on a multitude of datasets. Wellington outperforms other methods in almost all situations. An open-source software package, pyDNase, that facilitates interacting with DNase-seq data and provides many tools for DNase-seq analysis is also presented. Wellington is used to perform footprinting on clinical samples to validate cell lines as a model system, and to identify the binding partners of the RUNX1/ETO fusion protein in t(8;21) AML. By expanding the Wellington method, differential footprinting is shown to be able to link differences in transcription factor binding at promoters to changes in gene expression. Applying this methodology to a range of haematopoietic cell types illustrates the ability for differential footprinting to identify key regulators in the haematopoietic lineage. These results represent advances in the methods available to analyse DNase-seq data (all of which have been released as free, open-source software) and demonstrate the power of integrating DNase-seq footprinting with other functional genomic assays to study transcriptional regulation.

Abbreviations and Acronyms

-seq	sequencing
AML	Acule Myeloid Leukaemia
ATAC	Assay for transposase-accessible chromatin
AUROC	Area Under the Receiving Operating Characteristic
ChIP	Chromatin Immunoprecipitation
DHS	DNase Hypersensitive Site
DNA	Deoxyribonucleic Acid
dsDNA	double stranded DNA
DNase	Deoxyribonuclease
ENCODE	Encyclopedia of DNA Elements
FACS	Fluorescence Activated Cell Sorting
FAIRE	Formaldehyde Assisted Isolation of Regulatory Elements
HAT	Histone Acetyltransferase
HDAC	Histone Deacetylase
PCR	Polymerase Chain Reaction
PIC	(Transcription) Preinitiation Complex
PWM	Position Weight Matrix
PSSM	Position Specific Scoring Matrix
ROC	Receiving Operating Characteristic
SDS-PAGE	Sodium dodecyl sulfate - Polyacrylamide gel electrophoresis

Chapter 1

Introduction

1.1 Motivation

Understanding the information encoded in the genome requires the study of how transcription factors can recognise DNA sequences in order to coordinate gene expression. The motivation behind the work presented in this thesis stems from the need for a high-throughput assay to identify occupied transcription factor binding sites without the use of antibodies, as used in chromatin immunoprecipitation (ChIP) experiments, which are currently considered the 'gold standard' for identifying protein-DNA interactions. In this thesis, I will introduce digital DNase I footprinting, and illustrate how it can be used as a complementary method to genome-wide ChIP experiments (ChIP-seq) for the global identification of occupied transcription factor binding sites. At the onset of this PhD project, digital footprinting after DNase-seq[1] was a novel technique, having only been described two years earlier [1]. Due to the unique challenges presented by the data, there were only a handful of laboratories performing DNase-seq and even fewer with the necessary experience to analyse the data. Here I sought to develop software to facilitate the analysis of high-read depth DNase-seq data with the aim to identify DNA-sequences occupied by transcription factors, and evaluate the extent to which DNase-seq is an accurate assay for determining transcription factor binding sites by benchmarking current analysis methods and other transcription factor binding assays alongside these novel analyses.

[1]Not to be confused with DNA-seq (DNA Sequencing).

1.2 Overview

This thesis is presented in the format of a thesis based on publications, with a brief introduction and discussion. Each results chapter is composed of novel research that has either been submitted or accepted for publication in a peer-reviewed journal. Each of these chapters are prefaced with a statement outlining the motivation for, and overview of, the research, and personal contributions towards the work performed.

The *Introduction* provides the motivation for the work conducted and concisely introduces methodological aspects of the thesis that are used extensively but not described in subsequent chapters: classical methods for identifying transcription factor binding sites, chromatin immunoprecipitation, DNase footprinting, and motif finding.

The first paper [2], referred to as the *Wellington paper*, extends the introduction by providing a comprehensive overview and critical analysis of current methods of analysing DNase-seq data. Here, a novel method to identify transcription factor binding sites from DNase-seq data, called Wellington, is described. Wellington is objectively benchmarked against both ChIP-seq and previous algorithms designed for the analysis of DNase-seq data, showing that Wellington outperforms previously described methods across almost all performance metrics. Alongside Wellington, pyDNase, a software package to facilitate the analysis of DNase-seq data, is also described.

The second paper, referred to as the *differential footprinting paper*, advances the Wellington method developed in the *Wellington paper* to allow the comparison of two datasets in order to identify differentially occupied protein binding sites between two DNase-seq datasets. A comparative analysis of DNase-seq experiments with cells from healthy donors over a range of cell types illustrates the possibility of using differential footprinting to inform gene expression prediction models. This paper also describes several improvements to the underlying pyDNase software library, including increases in speed, new analysis scripts, the ability to correct for DNase I cutting bias when visualising data, and a DNase-seq analysis tutorial for those new to DNase-seq analysis and footprinting.

A summary of the work alongside an outlook on DNase-seq footprinting and conclusions of the thesis are present in the *Discussion*.

1.3 Background

1.3.1 Transcriptional regulation

Transcriptional regulation is one of the many underlying mechanisms by which the cellular state can be modulated. In single-celled organisms, genes are switched on and off in response to changing levels of nutrients and other intra- and extracellular cues. While the same is true for multicellular organisms, transcriptional regulation is also pivotal to the ability to utilise a single genome in order to generate a multitude of cell types. This cellular differentiation is driven by tissue-specific patterns of gene expression that are guided by spatial, temporal, and environmental cues throughout development during the lifetime of an organism [3].Although it may be biologically feasible for protein levels to be completely regulated at the post-translational level, this is not the case, and transcriptional regulation underpins genetics and is fundamental to all life.

The correct tissue-specific and temporal function of the genome is tightly controlled by transcription factors, proteins that bind specific DNA sequences in order to regulate gene expression. Upon binding, transcription factors can recruit other proteins to modulate gene expression or alter chromatin architecture. In the most basic example, a transcription factor binds directly adjacent to the transcriptional start site (TSS) of a gene (the promoter), recruiting the transcription pre-initiation complex (PIC), which is responsible for the positioning of RNA polymerase II at the TSS, leading to the transcription of the gene into RNA, thus promoting gene expression [4, 5]. Conversely, transcriptional repressors can down-regulate transcription by various mechanisms such as inhibiting the assembly of PIC, or preventing other promoting transcription factors from binding [6].

1.3.2 The chromatin landscape

The ability for a transcription factor to promote gene expression is conditional on the ability of the transcription factor to physically contact (bind) the DNA. In eukaryotes, the majority of nuclear DNA is bound to histone proteins and as a consequence is transcriptionally silent [7, 8]. Heterooctameric histone complexes are encircled by 147bp of DNA, resembling a thread around a spool — a complex known as a 'nucleosome.' Nucleosomes form interactions with scaffolding proteins (the most abundant of which is histone H1) in order to form compact chromatin fibres [9] (Figure 1.1).

Chromatin serves two functions: it allows the vast amounts of DNA present in

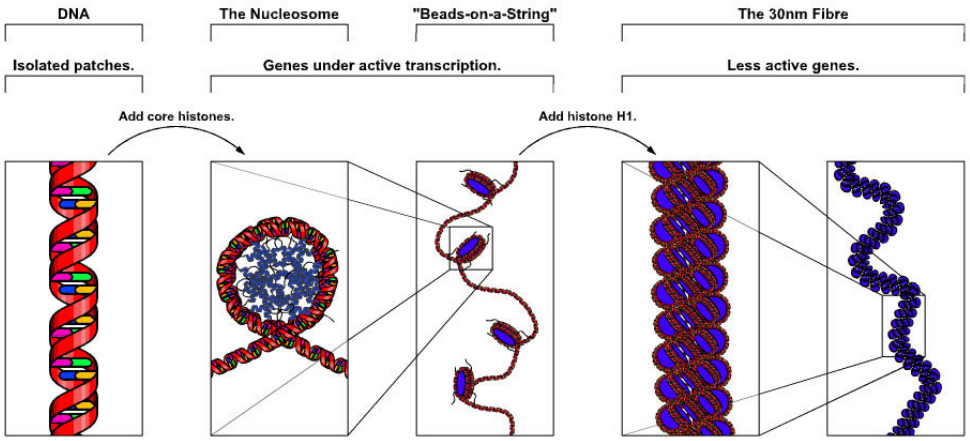


Figure 1.1: **The role of nucleosomes in chromatin packing.** A diagrammatic representation of naked genomic DNA, a single nucleosome, loosely packed euchromatin, and a densely packed 30nm heterochromatin fibre. Adapted from [10].

the nucleus of each cell to fit within the confines of the nucleus, and provides an additional layer of genetic regulation. Chromatin is often described of as being in two states — 'open' and 'closed' (or 'accessible' and 'inaccessible' to transcription factors). Consistency or alteration of covalent modifications of the N-terminal histone protein tails leads to the maintenance of the current chromatin state, or a remodelling cascade that can lead to nucleosome repositioning or removal (depletion) by chromatin remodelling complexes [11]. These histone modifications themselves can be driven by transcription factors that indirectly promote the opening of chromatin (pioneering factors) [12], close the chromatin (repressors), or maintain the current chromatin state. The ability of a transcription factor to bind a consensus sequence and promote gene expression is therefore often limited by the accessibility of the DNA to the transcription factor due to chromatin packing — in any given sample of human cells, roughly 90% of the genome is 'closed chromatin.' A good review of chromatin architecture in the context of DNase Hypersensitivity is provided in [13].

1.3.3 Regulation from a distance

In eukaryotes, the transcription of genes is not only regulated via their promoter, but by distal enhancers as well. Enhancers are regulatory regions of DNA that regulate gene transcription from a distance (i.e. more than 2kb from a gene's TSS) via

interactions with one or many gene promoters, as discovered in some cases [14, 15]. Conservative estimates place the number of enhancers in the human genome at $> 100,000$ [16]. Liberal estimates place this as high as $2,900,000$, which vastly outnumbers the ca. $25,000$ protein coding genes in the human genome. Enhancer activity is highly cell type specific [17], with only around $200,000$ enhancers being active at any one time [18]. Enhancers, like promoters, are also bound by transcription factors. They function via the induction of chromatin looping, making contact with and activating one or several distal promoters [19]. Conversely, enhancers can also be bound by silencing factors that can inhibit the transcription of genes by various mechanisms. Neuron-Restrictive Silencer Factor (NRSF), for example, inhibits the expression of neuronal specific genes in non-neuronal tissues by binding to enhancers with the Neuron-Restrictive Silencer Element (NRSE) sequence, recruiting histone deacetylases (HDACs) that lead to histone hypomethylation and inducing chromatin packing, silencing the enhancer [20]. The most abundant transcriptional repressor, CTC-binding factor (CTCF), however, inhibits the action of transcription factors via a different mechanism, inhibiting chromatin looping preventing the contact of nearby enhancers to their target promoters [21]. This illustrates the complex many-to-one relationship of enhancer effects on gene expression, whereby gene promoters integrate the signals from many gene-promoting and gene-repressing enhancers, themselves only active in specific contexts, in order to regulate gene expression.

1.3.4 Characterising regulatory elements

The publication of the finished sequence by the human Genome Project in 2004 [22] was a milestone in genomics, however, much work was still required to identify all the functional elements of the genome such as protein coding and non-protein code genes, alongside regulatory elements (also known as functional non-coding regions) such as promoters and enhancers. The US National Human Genome Research Institute (NHGRI) launched the Encyclopedia of DNA Elements (ENCODE) project consortium with the aim of annotating functional elements in the human genome over multiple cell lines [2]. The results from the pilot study on just 30Mb (1%) yielded a wealth of results, which not only recapitulated known methods of genetic regulation that had not previously been characterised on a systems level, but also revealed novel information [23]. One example is the finding

[2]The biggest criticism of the ENCODE project is its decision to use cell lines and not healthy samples in the majority of their assays.

that specific transcription factors previously thought to only bind gene promoters were found to bind to enhancers, and regulatory elements previously annotated as enhancers were discovered to be novel promoters for unannotated transcripts.

Discovering promoters is a relatively simple task achieved by identifying the region directly adjacent to a transcriptional start site. Unlike protein coding genes, there are no sequence-specific elements that accurately predict the presence of an enhancer although some sequence features, like increased CG content and sequence conservation are highly correlative. As part of the full ENCODE project, the ENCODE project consortium employed several high-throughput genomic techniques in an attempt to identify and characterise regulatory elements, including the identification of transcripts through CAGE-seq [24] and RNA-seq [25], transcription factor binding sites via ChIP-seq and DNase-seq, and the determination of chromatin structure via ChIP-seq, MNase-seq [26], FAIRE-seq [27], and DNase-seq. Whilst these assays have helped identify putative enhancers and distal regulatory elements, the knowledge of their existence by themselves does not provide characterise the action of these regulatory elements.

The data and analyses from the ENCODE project has been pivotal in facilitating our understanding of the genome, in particular, provided a wealth of information characterising the location of the non-coding regulatory elements in the human genome. However, the characterisation of these regulatory elements in the human genome has proven to be challenging. Given the approximately 1500 transcription factors, and the large number of enhancers that are estimated to be active in any one cell type at any given time [28], the functional characterisation of all of these elements is a monumental task that will require many different assays and analyses, and ENCODE has only scratched the surface.

Even with the knowledge of the location and the known transcription factor binding sites in an enhancer, it remains difficult to predict the genes are under the control of a specific transcription factor. There are, however, certain histone modifications (H3K27ac and H3K4me1) that are associated with active and inactive enhancers, along with specific co-activators (p300) that have been found to bind to active enhancers. It is thought that by understanding the specific combinations of transcription factors that bind regulatory elements, an 'enhancer grammar' can be inferred that enables the prediction of enhancers *in silico*.

The ENCODE project has provided a wealth of data about regulatory elements that have illustrated the importance of enhancers in evolution. It has been known that GWAS SNPs are enriched in non-coding regions, and with the characterisation of regulatory elements through ENCODE, it has subsequently been shown

that these SNPs are enriched in regulatory elements, or are in linkage disequilibrium with those elements [29, 30]. Moreover, evidence suggests that evolution is accelerated at non-coding regulatory elements [31]. Regulatory elements offer efficient ways to make phenotypic changes to an organism through only a small number of nucleotide polymorphisms which change the regulatory phenotype of a gene. In one such example of evolution through regulatory elements, Single Nucleotide Polymorphisms (SNPs) were identified via quantitative trait loci (QTL) analysis that were associated with the absence of pelvic fins in a specific subpopulation of sticklebacks lost through divergent selection; Transfecting embryos with an edited enhancer led to the reestablishment of pelvic fins [32], reinstating phenotypes that had been lost through selection. Individual SNPs in enhancers that are sufficient to disrupt a single transcription factor binding site have been shown to alter chromatin structure, causing the chromatin to close and prevent activation of the enhancer [33].

However, the types of assays performed by the ENCODE project that characterise the location, and in some cases, the transcription factor binding sites, do not identify where in the genome an enhancer is acting. Whilst expression quantitative trait loci (eQTL) is able to correlate changes in gene expression to regulatory elements these studies require very large samples in order to gain statistical power and it is impossible to show correlation vs causation. In addition, eQTL studies have been shown to be highly cell-type specific, which is probably due to the cell type specificity of enhancers leading to variants in regulatory regions only affecting those cell-types in which the enhancer is active [34]. Beyond computational and statistical methods, there are several high-throughput biological methods for identifying long-range interactions in a genome such as contact between enhancers and promoters. These methods rely on cross-linking cells with formaldehyde in order to freeze the cell in time. The DNA is then digested with a restriction enzyme in order to generate globules of cross-linked protein with strands of DNA intertwined. A variety of different techniques are then used to ligate DNA that were spatially near to each other in the cell, which can then be sequenced in order to determine the chromatin interaction landscape (Figure 1.2).

These '3C' based techniques include chromosome conformation capture (3C) [36], circularised chromosome conformation capture (4C) [37], chromosome conformation carbon copy capture (5C) [38], Hi-C [39], and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) [40]. Whilst these tools give an idea of the long-range interactions of enhancers and promoters, they do not help elucidate the precise transcription factors that are ultimately the driving forces

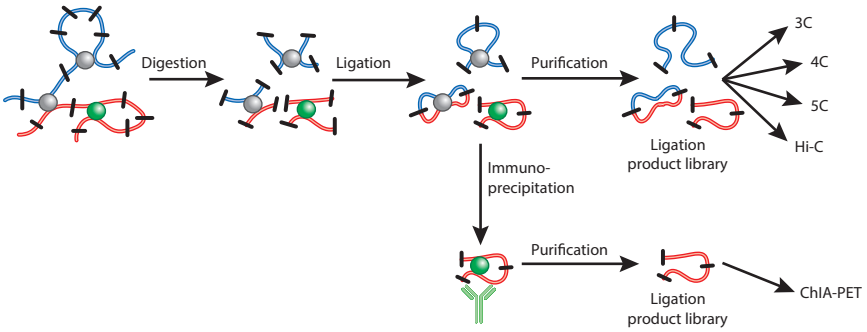


Figure 1.2: **Methods for identifying higher-order chromatin interactions.** Various methods exist for identifying long-range chromatin interactions. These '3C' based techniques cross-link chromatin with formaldehyde in order to isolate regions of DNA that are co-located in physical space. These regions are then circularised and sequenced, where each half of the sequencing read originates from one genomics location in contact with the other half. Image adapted from [35].

behind the enhancer/promoter contact. In order to piece together a system-wide view of gene regulation — the identification of active enhancers and characterisation of how these enhancers function in order to regulate their target gene(s) — a multitude of assays must be integrated. Information on transcription factor binding sites, long-range chromatin interactions, gene expression, histone modifications, mutation, genetic variants, amongst others, need to be analysed in concert in order to construct a complete model of transcriptional regulation in the human genome.

1.4 Identifying transcription factor binding sites

1.4.1 Chromatin immunoprecipitation

Currently, the gold-standard high-throughput method for identifying transcription factor binding sites is ChIP-seq. Almost all methods that aim to predict transcription factor binding sites have used ChIP-seq recapitulation as the validation metric. Both of the results chapters in this thesis will refer to ChIP-seq data, so a brief overview of the protocol, along with comments on the benefits and limitations, are outlined here.

The first stage in a ChIP-seq experiment is the cross-linking stage. Formaldehyde is added to live cells, reacting with the protein in the cell and resulting in neighbouring proteins becoming cross-linked through the formation of methylene bridges. DNA becomes trapped in this matrix of cross-linked proteins, providing a snapshot of the protein-DNA interactions in a population of cells. Cells are lysed, and the protein-DNA matrix is extracted and sonicated in order split the mixture into protein-bound fragments of DNA between 100 and 300 base pairs in length. An antibody for a protein of interest is then used together with a solid matrix to which it binds to enrich fragments of DNA bound by the target protein. The cross-linking process can be easily reversed by heating in water, yielding fragments of DNA that were bound to the protein of interest, either directly or indirectly. These fragments are then analysed by high-throughput sequencing, and are mapped to reference genome of the organism being studied in order to determine the genome-wide binding sites of the protein of interest [41] (Figure 1.3).

Analysing ChIP-seq data allows for the identification of putative binding sites with a resolution in the range of 400—2000bp. Because of this, it can extremely difficult to identify whether the protein of interest is directly or indirectly binding a given region unless further experiments are carried out, and results from ChIP-seq experiments can yield numerous false positives. Therefore, ChIP-seq experiments are often required to be validated through other experimental means. One such method to address this issue is ChIP-exo [43] (Figure 1.4). An exonuclease is used to trim the overhanging fragments surrounding a transcription factor complex. This method has not yet gained widespread adoption, with only a handful of publicly available datasets (43 as of 13th December 2014 contrasted with 11,443 ChIP-seq datasets) in the Gene Expression Omnibus [44, 45].

The greatest limitation of ChIP-seq (and therefore ChIP-exo) stems from the

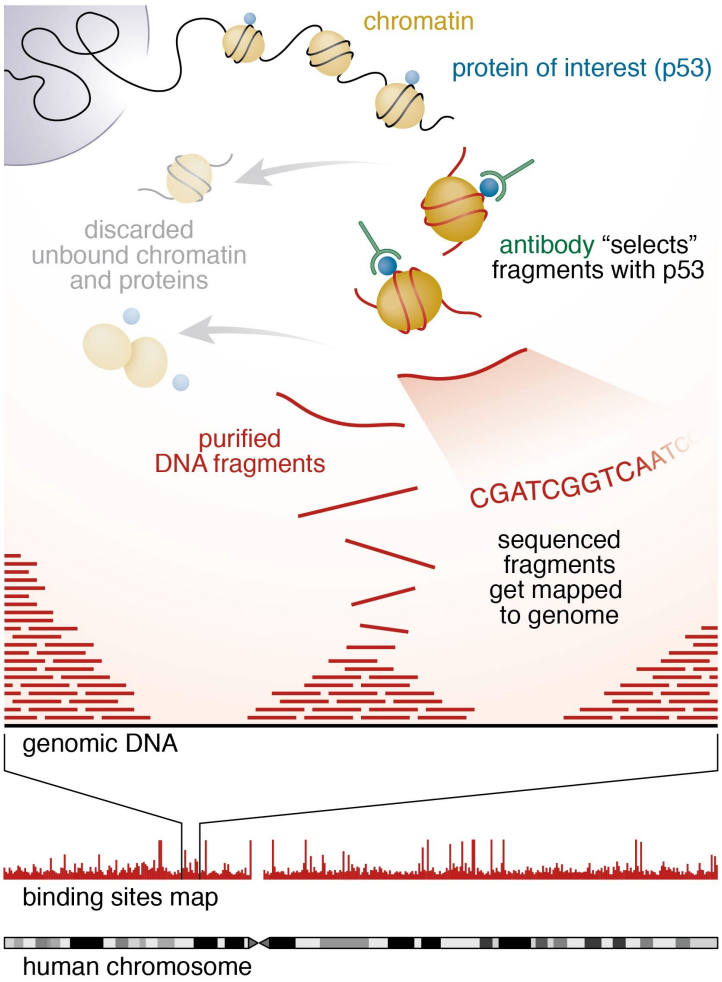


Figure 1.3: **Chromatin Immunoprecipitation.** Cells are cross-linked using formaldehyde, lysed, and sonicated. An antibody against a protein of interest is subsequently used in order to isolate regions of DNA associated with a specific protein. Next, the cross-linking is reversed so that the DNA associated with the protein-DNA interaction can be sequenced and aligned to the genome in order to identify transcription factor binding sites. Image courtesy [42].

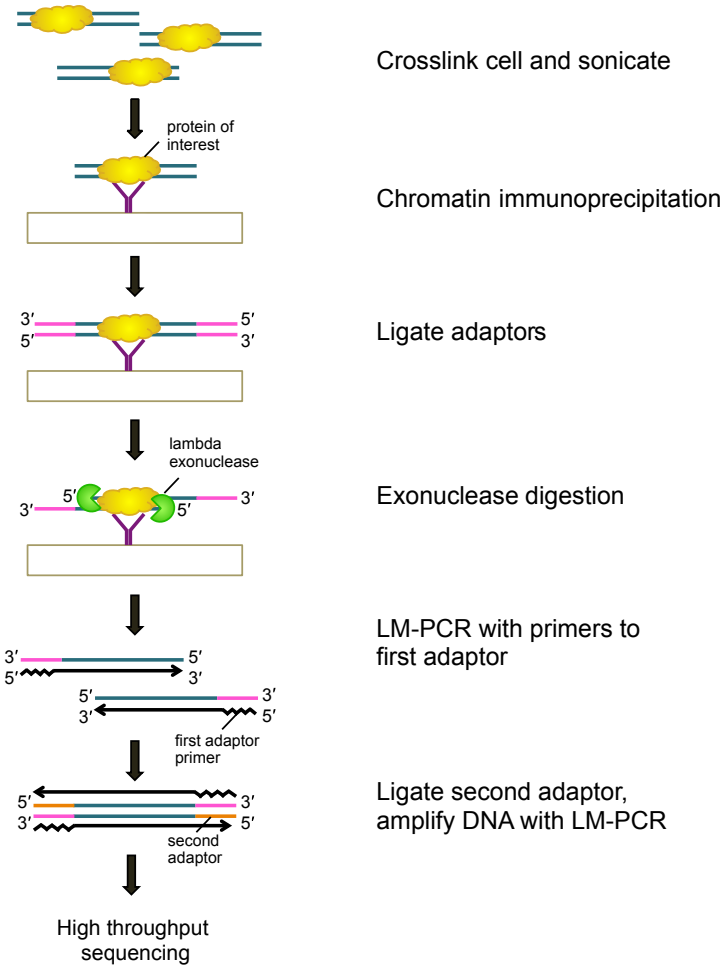


Figure 1.4: **Chip-exo.** A standard ChIP protocol is followed, however, before elution, the 5′ end of the bound DNA fragment is trimmed using lambda exonuclease, so the 5′ end of the DNA flanks the protein-DNA interaction. These 5′ ends can then be targeted for sequencing using LM-PCR. Image courtesy [46].

requirement that a highly specific, high-affinity antibody must be raised against the protein of interest. This limits experiments to one protein per sequencing run, as well as to proteins that are already of interest and known. The inability to distinguish between direct and indirect binding to the DNA is another major limitation of ChIP-seq. The absence of the target transcription factors corresponding to DNA binding motifs provides evidence of indirect binding, however the mere presence of a motif is not a good indicator that the protein is directly bound. In addition, epitope masking can occur, in which the antibody used in the immunoprecipitation is unable to recognise the target protein. Solutions such as epitope tagging with an accessory protein (e.g. GFP) can be utilised to mitigate this problem. Epitope tagging can also allow the study of proteins for which no antibody is available, but this could interfere with the cellular function of the epitope tagged protein (a review highlighting the technical considerations for ChIP-seq can be found [47]).

1.4.2 DNase I footprinting

The identification of protein-DNA interactions via DNase I began with the observation that upon ultraviolet irradiation, the breakage of DNA of the *E. coli lac* operator was diminished at specific sites by the presence of the *lac* repressor [48], and similarly, the ability of dimethylsulfate (DMS) to methylate purines (preferentially guanines) was also inhibited by the presence of a bound protein interacting with these bases [49]. The ability of a DNA binding protein to shield the DNA from damage was harnessed in a technique called deoxyribonculease I (DNase I [3]) footprinting [50]. DNase is a eukaryotic endonuclease that cleaves the DNA phosphodiester backbone adjacent to pyrimidic bases on one strand at a time i.e. introduces 'nicks'. It has an observable sequence specificity [51–54], where the phosphodiester bonds adjacent to specific sequences are hydrolysed with preference over several orders of magnitude. However, even though this bias exists, it is unlike a bacterial restriction enzyme in that it does not have a recognition site.

In the original DNase footprinting protocol, by subjecting a protein-DNA complex to DNase I cleavage, a southern blot was be used to visualise where the DNA is bound by a protein. DNase is unable to cleave DNA where protein is bound, but *is* able to cleave directly adjacent (as much as steric hindrance allows) to the protein-DNA complex and anywhere else where the DNA is not bound to a pro-

[3]Interestingly, DNase was originally referred to as 'DNAase'[50] and has at at some point lost the superfluous 'a.'

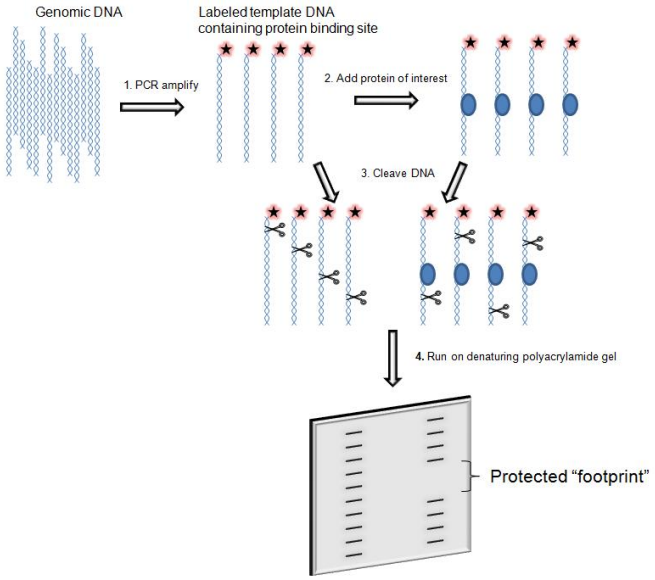


Figure 1.5: **DNase I footprinting via southern blotting.** Protein-DNA interactions can be identified *in vitro* using DNase I footprinting. A protein-DNA complex is subject to digestion by DNase I. The DNA fragments generated by this process are subject electrophoretic separation, where a 'footprint' is left by the protein contacting the DNA and preventing DNase I digestion. Image courtesy [55].

tein. This method was coupled with DNA sequencing (at the time of discovery, Maxam-Gilbert sequencing) in order to identify the sequence of the protein-DNA binding site. DNase footprinting provided conclusive evidence that transcription factors bind specific DNA sites based on their sequence (Figure 1.5).

DNase footprinting was quickly adapted for *in vivo* use via ligation-mediated PCR (LM-PCR) [56]. Whether whole cells or isolated nuclei are subjected to DNase digestion, and regions of interest are amplified via single-stranded PCR using primers flanking the region of interest. As nicks introduced by DNase will prevent elongation by DNA polymerase, the PCR fragments will abruptly terminate in regions where DNase is able to cleave the phosphodiester backbone. In regions where a protein is bound, fewer nicks will be present and the PCR reaction will rarely terminate in these regions. Linkers are then ligated to the single-stranded DNA fragments and primers against these linkers utilised to reconstitute the complementary strand of DNA. This ability to characterise the activity of transcription factors in their native environment is critical to understanding genetic regulation. DNase footprinting methodology remained largely unchanged as a single locus assay for 30 years, and was restricted to the study of limited numbers of proteins and sequences at once due to the laborious nature of

the assay. During this time, in terms of the number of experiments submitted to the main online data repositories, the Gene Expression Omnibus (GEO) and the Sequence Read Archive (SRA), ChIP-seq remained the most popular assay for the determination of transcription factor binding sites.

High-throughput DNase I assays

In 2004, two member groups [57, 58] of the ENCODE consortium developed renewed interest in DNase as a tool to identify regulatory regions by using the cleavage activity of DNase to map thousands of so-called DNase Hypersensitive Sites (DHSs). In order for regulatory regions to function, nucleosomes are displaced, revealing regions of 'open' chromatin whereby the DNA becomes accessible to transcription factors. These regions, often 200-2000bp in length, are much more sensitive to cleavage by DNase than 'closed' chromatin i.e. DNA that is bound to nucleosomes. The authors showed that short fragments of DNA isolated from DNase cleavage reactions aligned to known regulatory regions in the human genome, namely known enhancers and promoters .

Shortly afterward, the use of DNA microarrays was employed [60] in order to provide the first high-throughput screen for DHSs, and with the increased affordability of next-generation high-throughput sequencing technologies, DNase digestion was coupled with high-throughput sequencing (DNase-seq), providing the first unbiased assay for identifying active regulatory regions [1, 60]. The specific method used for library preparation was introduced as the 'single hit' method, Figure 1.6, with the slight disadvantage that sequence fragments generated by the single-hit method are limited to 20bp, which although theoretically mappable, is at the lower limit of acceptable read length to be able to align to the human genome.

1.4.3 DNase-seq for identifying transcription factor binding sites

Unlike the work undertaken by Galas and Schmitz [50], the sequencing depth of these assays was insufficient to distinguish areas of protection from DNase cleavage by bound proteins, but the method identified potential regulatory elements in the human genome. It was not long, however, before this method was adapted as a high-throughput method to identify DNase I footprints *in vivo*. High-throughput sequencing platforms typically sequence 20-72bp of a ca. 300bp fragment in the $5'$ to $3'$ direction. When aligning the sequenced fragment to the reference genome,

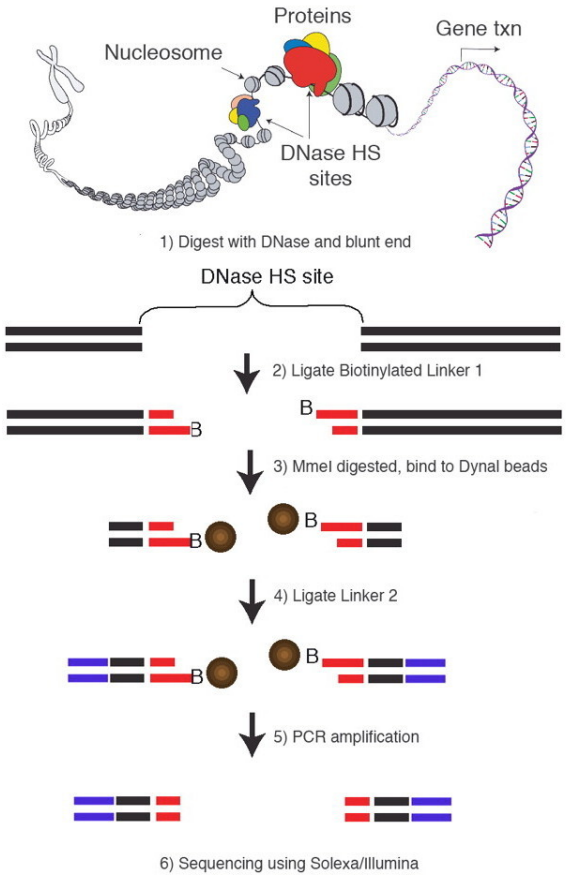


Figure 1.6: **The 'single-hit' DNase-seq method**. After DNase digestion, fragments are ligated to a linker which contains a MmeI restriction site. The sample is treated with MmeI, which will cut 20bp downstream from the recognition site, and finally ligate this fragment to another linker to 'pad' the fragment to be sufficiently long enough to be sequenced. Adapted from [59].

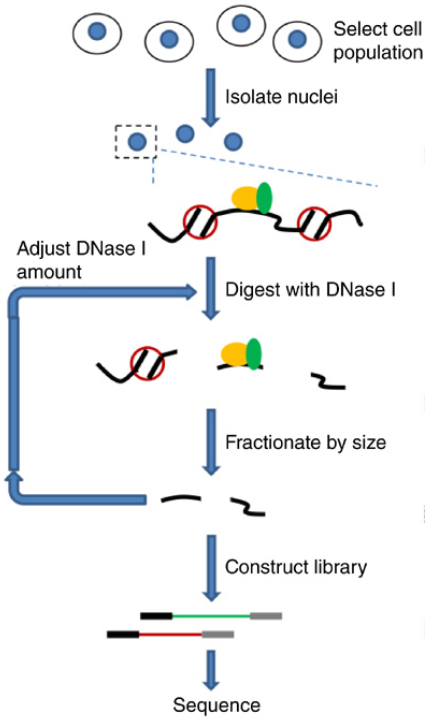


Figure 1.7: **The 'double-hit' DNase-seq method**. Cells are subjected to DNase I digestion *in vivo*, where DNase I is only able to digest DNA in regions of open chromatin where the DNA is unbound by protein. Fragments generated by this process that pass a size-selection threshold are isolated and then sequenced. Adapted from [62].

the 5′ end of the aligned sequence corresponds to where the DNA has been cut by the DNase, and the 3′ end corresponds to the sequence length limit of the sequencer. By aligning reads to the genome and identifying the location of all the 5′ ends of the sequenced fragments, the genomic position where DNase has cut DNA can be established.

Yeast was chosen as the first organism for this assay, which the authors named 'digital genomic footprinting' [61] for two main reasons. Yeast's small genome (12Mb) yields 400× the sequencing depth compared to the same number of sequencing reads on a human sample, and the transcription factor binding sites within the yeast genome are well classified, providing a simple method of validating the results from this novel technique. Although these experiments did not provide much in the way of elucidating any novel biological results, they provided proof of principle that DNase digestion followed by high-throughput sequencing could identify transcription factor binding sites *in vivo*.

Further developments drastically simplified the DNase-seq protocol. In the

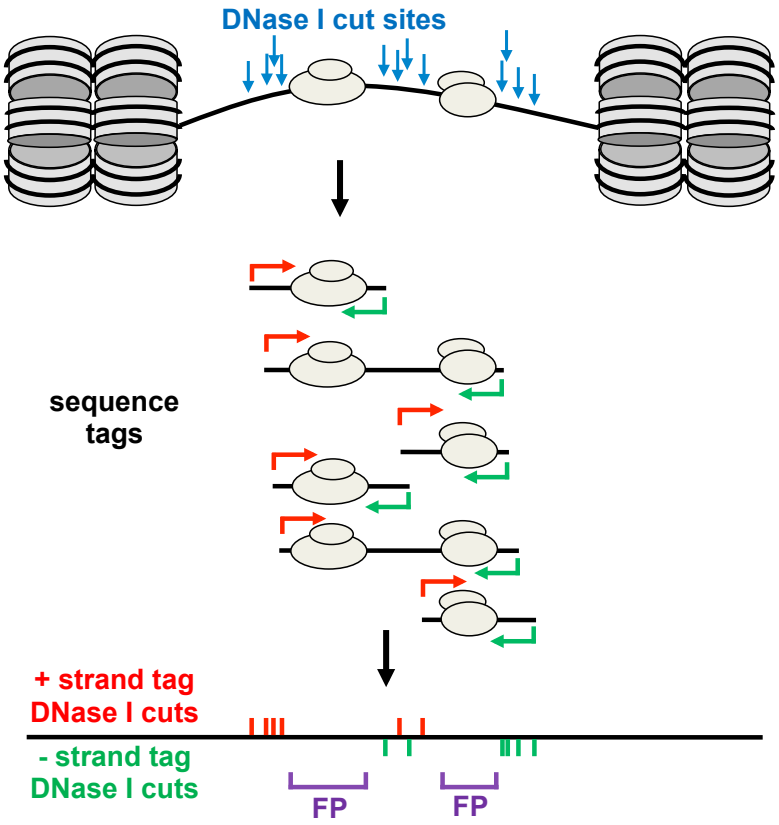


Figure 1.8: **DNase-seq Footprinting**. Classical DNase I footprinting can be adapted into a whole-genome study by using the 5′ end of aligned sequence tags as a measure of DNase I cutting. By subjecting cells to digestion by DNase I and isolating sub-nucleosomal fragments (<150bp), the 5′ most end of aligned sequence tags will occur in protein-DNA binding sites less often than of neighbouring regions due to protection of the DNA from DNase I cleavage.

'double-hit' protocol [63], the ligation of MmeI linkers and digestion was replaced by a simple size selection step (Figure 1.7). Regions of open chromatin are digested with DNase I at the correct concentration that provides two 'cuts' per DNase hypersensitive site. The fragments that span the protein-DNA binding sites are isolated using size selection on an agarose gel or via ultracentrifugation using a sucrose column. These data can can be analysed in order to detect protein-DNA binding events (footprints) (Figure 1.8).

The assay is extremely sensitive to the precise distribution of fragment sizes, with fragments less than 50bp resulting in fragments that span the protein-DNA binding site, which is best suited for footprinting [52]. Fragments larger than 150bp often span entire nucleosomes, and these larger fragments are not suitable for identifying transcription-factor binding sites. Using either a combination of

size-selection, or *in silico* size selection following paired-end sequencing, DNase-seq can be used not only to find transcription-factor binding sites, but also nucleosome positions [64]. Over 90% of DNase-seq data in the public domain (almost all of which is part of the ENCODE project) have been generated with this simpler 'double-hit' method, with only the very early experiments using the older single-hit method.

1.5 Analysing transcription factor binding data

1.5.1 Peak calling

The basic analysis of both ChIP-seq and DNase-seq data is to identify where in the genome the protein of interest is binding ,or in the case of DNase-seq, the identification of DHSs. The aligned sequence tags are analysed to find regions of the genome where the number of reads that have been aligned is statistically significantly higher compared to a background model, whereby the exact statistical methods used vary between the tools used to identify the peaks. These binary predictions partition the genome into regions where the protein is bound to the DNA (peaks) and regions where it is not.[4] When analysing ChIP-seq data, the scores assigned to these regions are often used as evidence of the 'strength' of binding. However, as there is no easy method of determining the heterogeneity of transcription factor binding in the population of cells used in the assay, it is not possible to differentiate between a protein binding weakly in all cells in the population, or binding strongly in a subpopulation. A wide variety of tools have been developed to fulfil the purpose of peak calling, and the sophistication of these methods ranges from using a simple sliding window to calculate the sum of reads in an area, to methods that consider the signal on a single base-pair resolution, Hidden Markov Model based approaches, and those that correct biases in the underlying library preparation protocols and sequencing methods used to generate the data. Much could be said about the approaches taken to analyse ChIP-seq data, with different ChIP-seq peak callers yielding peaks that drastically alter the biological interpretations of the downstream analyses (reviewed extensively in [65]).

[4]There is much to be said about such a binary view on the protein-DNA landscape, whereby near to this threshold, the difference of one aligned read can determine the difference between 'bound' and 'unbound'.

1.5.2 Digital genomic footprinting

For ChIP-seq analyses, identifying regions of enrichment (peaks) followed by motif searching (outlined in Section 1.5.3) is the preferred analysis in order to identify the transcription factor binding sites. Similar analyses can be performed on DNase-seq data through the identification of motifs in DHSs, but there can be between 2—100× the number of DHSs in a DNase-seq experiment compared to a ChIP-seq experiment, with the number of DHSs per cell line ranging between $84,201$ for Th1, $142,986$ for K562, and $266,618$ H7-hESC [66], whereas the number of ChIP-seq peaks ranging between $1,902$ for NRF1 and for $45,732$ CTCF in K562 cells [2]. Because of the large numbers of DHSs detected, the number of transcription factor binding DNA motifs present in these regions will far outnumber the number of transcription factors that are actually bound, so this analysis can yield many false positives. The power of DNase-seq is the ability to analyse the data in order to demarcate transcription factor binding sites within DHSs (i.e. footprinting). This can performed computationally by identifying short (< 50bp) regions where the number of $5'$ sequence tags aligning to the genome is diminished due to a protein-DNA interaction blocking digestion of the DNA by DNase. This process, originally referred to as 'digital genomic footprinting' was first described using data generated on the single-hit data in yeast, by searching for regions of depleted $5'$ sequence tags in the data using a binomial test [61].

The first DNase-seq footprinting efforts on the human genome were performed using the early 'single-hit' data generated by the ENCODE project [67, 68]. These methods used fundamentally different statistical analyses such as a Bayesian Hierarchical Model and a Hidden Markov Model, respectively, to identify footprints. DNase-seq data from 41 cell lines generated as part of the ENCODE project provided the first comprehensive overview of transcription factor footprints in the human genome [66]. The authors also described their approach a novel algorithm (named *Ambrose*) for discovering footprints without motif information, along with a set of footprints for 41 cell lines with ChIP-seq validation for a select few transcription factors in K562 cancer cell line (discussed further in Chapter 2).

In contrast to ChIP-seq, DNase-seq footprinting provides a holistic approach to identifying transcription factor binding sites. DNase-seq reveals that a specific 10-30bp fragment of DNA is bound, without identifying which protein is bound. ChIP-seq determines which protein is binding a certain region, but not where it is bound, and whether it is binding directly or indirectly. Because of this, ChIP-seq and DNase-seq have complementary roles, and the combination of ChIP-seq and

DNase-seq can be used to differentiate primary from secondary binding [66]. Several studies have shown that transcription factor footprints are able to recapitulate the binding of almost all transcription factors, as well as the identification of novel transcription factors. Several studies have commented on the false positive rate of these analyses, commenting on both the lack of evolutionary conservation of so-called 'novel' transcription factor binding sites [2, 69], and the possibility that these false positives arise from unaccounted bias in DNase I cutting [52].

A thorough introduction to DNase-seq footprinting methodology is presented in Chapter 2.

1.5.3 Identifying DNA binding motifs

After peak (or footprint) detection, the most fundamental downstream analysis of transcription factor binding sites as determined by ChIP-seq is the identification of the DNA sequence that a transcription factor binds to by analysing the set of sequences identified by the peak calling process. Knowledge of the DNA sequences that are preferentially bound by individual transcription factors (binding motifs) allows the prediction of transcription factor binding sites within the genome, and supplements other genomic assays by being able to predict bound transcription factors based on DNA sequence (e.g. the determination of binding partners or the identification of common DNA motifs in gene promoters that may suggest common regulatory mechanisms). As most transcription factors exhibit degeneracy in their ability to recognise DNA sequences, to represent the complete repertoire of sequences that a transcription factor binds to, the position weight matrix (PWM), also known as a position-specific scoring matrix (PSSM)[70] is frequently used.

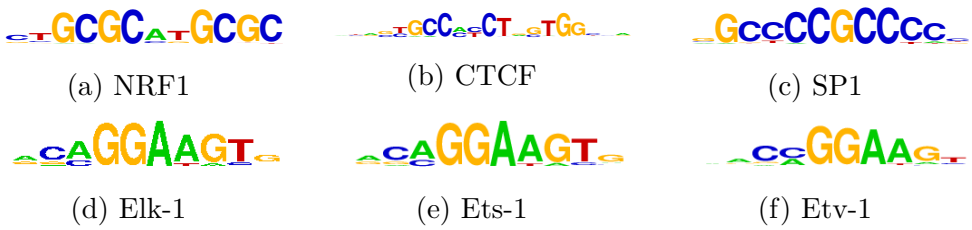
For a set S of N sequences of length l that a transcription factor is known to bind to, the elements of the PWM P are calculated as follows.

$$P_{k,j} = \frac{1}{N} \sum_{i=1}^{N} I(S_{i,j}, k) \tag{1.1}$$

where $i \in (1, ...N)$, $j \in (1, ..., l)$, $k \in \{A, C, G, T\}$, and the identity function

$$I(a, k) = \begin{cases} 1 & \text{if } a = k \\ 0 & \text{if } a \neq k \end{cases}$$

PWMs are more intuitively visualised (Figure 1.9), where it is easily observed that families of transcription factors often have extremely similar PWMs (Figure 1.9d-f), reflecting the fact they share a common DNA binding domains. The use



(a) NRF1 (b) CTCF (c) SP1

(d) Elk-1 (e) Ets-1 (f) Etv-1

Figure 1.9: **Position Weight Matrices (PWM) can be represented graphically to compare and contrast the binding characteristics of transcription factors.** A graphical representation of the PWMs for three transcription factors containing different DNA binding domains highlights the difference DNA sequences bound (a,b,c), whereas three different transcription factors from the Ets family that that share similar binding domains have very similar PWMs (d,e,f) The height of the letters indicates the bits of information that the specific letter carries.

of the PWM to represent a protein's DNA binding affinity makes the assumption that the effects of a nucleotide on the binding affinity of a protein are independent of all other positions in the PWM, and is a simplistic method of representing a complex biophysical interaction. Several more complex models for representing DNA binding sites have been developed, most have not managed to significantly improve on the PWM, which also had the added benefit that is simple to interpret visually. There have been developments in the area of fitting the 'best' PWM for a transcription factor, identifying the short 8-10bp sequences out of the 200—1000bp ChIP-seq that the transcription factor recognises, and these various methods rely on a number of statistical frameworks, comprehensively reviewed in [71].

The identification of the enriched set of motifs in a ChIP-seq dataset often reveals more than the sequence that the immunoprecipitated transcription factor binds. As transcription factors do not bind on their own, but interact with other factors, secondary motifs found in ChIP-seq peaks provide insight into the co-operative transcription factor binding partners of the protein of interest, and therefore the protein-protein interactions of the target protein. The total number of known human DNA binding motifs currently numbers in the thousands, and several databases exist such as TRANSFAC [72], JASPAR [73] which have catalogued them and which are constantly updated as a result of substantial high-throughput screening studies. The amount of redundancy is large, as many transcription factors share common DNA binding domains, resulting in families of transcription factors, with often diverse functions, that recognise similar DNA binding motifs (Figure 1.9). One extreme example is the ETS factor family of transcription factors, with 27 members in the human genome that all bind the relatively simple

core motif of TTCC [74]. Coupled with the degeneracy of DNA binding proteins, this many-to-many relationship where a single DNA binding site can be bound by a myriad of transcription factors, and conversely a single transcription factor can bind many different sequences with varying specificities illustrates the extremely complex nature of the transcription factor landscape within the cell.

Because of the large number of DNA binding motifs, almost any given region of the genome will contain scores of known DNA binding motifs.

1.6 Introduction to the thesis

DNase-seq is a high-throughput adaptation of a the classical DNase I Footprinting assay [50] that can identify regions of open (active) chromatin. Through the use of more sophisticated analyses, DNase-seq footprinting is able to demarcate transcription factor binding sites with an accuracy of < 30bp resolution [61]. DNase-seq data continues to be generated by public functional genomics consortia such as the ENCODE and NIH Roadmap Epigenomics projects, and the application of DNase-seq on mice [75], *Drosophila* [76], *Arabidopsis* [77], Rhesus monkeys, and Chimpanzee [78] illustrate the versatility of the method. The DNase-seq method is of great utility in organisms where little is known about transcriptional regulation. In these cases, the power gained from DNase-seq footprinting is not recapitulating results, but being able to study multiple transcription factors in one assay without the requirement to either specify a protein of interest or obtain antibodies as used in ChIP.

Despite this growing adoption of the method, prior to the work described in this thesis, the options available for analysing DNase-seq data were sparse. Several tools designed for peak calling in ChIP-seq data have been coaxed into this role, with tools such as MACS, FindPeaks, and HOMER being used to locate regions of open chromatin, along with several DNase-seq specific peak callers specifically for DNase-seq data, reviewed in [79]. For more fine-grained footprinting analyses, even though a number of papers had been published either describing or utilising footprinting analyses of DNase-seq data, there was no software available for performing footprinting, the exception being CENTIPEDE which is not a *de novo* footprinting method as it requires prior knowledge of genomic DNA binding motifs for a transcription factor of interest (discussed further in Section 4.4).

The aim of this thesis is to build on the pioneering work performed on developing the DNase-seq protocol, and the early attempts at footprinting DNase-seq data. Chapter 2 reviews the methods available for identifying transcription factor

binding sites in DNase-seq data and identifies several shortcomings in the methods being used to validate these attempts. Here, Wellington, a novel algorithm for identifying transcription factor binding sites is introduced, and benchmarked against ChIP-seq data and the results of other algorithms. The strand imbalance inherent in the DNase-seq library preparation protocol is shown to increase the predictive power of transcription factor binding site predictions. pyDNase is also introduced alongside Wellington, providing an easy to use and efficient application programming interface (API) for interacting with DNase-seq data, along with several convenient scripts for performing common analyses.

Chapter 3 builds on the methodological and analytical work performed in the previous chapter. An extension to the Wellington algorithm is presented that allows for the ability to identify differential transcription factor binding in common DNase hypersensitive sites, along with significant computational performance gains to the original method brought about by major refactoring of the code base. The analysis of a number of samples from primary cell lines using this extension identifies transcription factors that convey cell identity across the haematopoietic lineage, and is able to illustrate how transcriptomic data can be linked to transcription factors that affect gene expression in these cells.

Chapter 2

Footprinting analysis of DNase-seq data

2.1 Motivation

At the beginning of this project, several approaches to identifying footprints in DNase-seq data had been described. One such method outlined the footprinting analysis of yeast DNase-seq data generated using the single-hit method, with the authors providing a software implementation as a collection of scripts in MATLAB, Python, and Bash [61]. Another approach developed on single-hit DNase-seq data, but on human data, was CENTIPEDE [67] (which requires the locations of DNA binding motifs in the genome *a priori*), implemented in R. Neither of these tools were intuitive to use, requiring the DNase-seq data in non-standardised file formats to produce footprint predictions. The authors of these studies provided descriptions of these file formats and example files, but software to prepare these files from the sequencing data was not provided. Two further methods existed for which there was no software implementation available [66, 68], but the results of the analyses on several cell lines had been published.

Here, a comprehensive evaluation of the performance of these approaches to identifying protein-DNA interactions in DNase-seq data on identical datasets was performed, the first objective assessment of these different methods. As part of this effort, a software library for interacting with the raw DNase-seq alignment data called pyDNase was developed, in order to allow for the fast and efficient generation of the non-standardised data formats required by CENTIPEDE [67] and the Hesselberth Method [61].

Using pyDNase, a novel footprinting algorithm, *Wellington*, has been intro-

duced, that utilises a previously undescribed feature in the 'double-hit' DNase-seq protocol in order to increase the predictive power of DNase-seq footprinting. Wellington was validated against ChIP-seq data, motif content, and phylogenetic conservation scores across several cell types and between the two prevailing library preparation protocols. Where data are available, Wellington was benchmarked against other methods.

Here, a comprehensive suite of benchmarks over several criteria is presented that illustrate the power of Wellington to accurately identify protein-DNA interactions from DNase-seq data. In almost all circumstances, Wellington outperforms previously described methods across almost all performance metrics. Wellington is provided as part of the pyDNase package, a free, open source, and easy-to-install Python library. pyDNase was designed using modern software development practises such as unit tests, extensive documentation, and continuous integration in order to maintain quality assurance throughout development of the tools.

2.2 Contributions

For this paper, I was responsible for the design of the study and the entirety of the design, development, and implementation of the underlying Wellington algorithm and pyDNase software, the data analysis, figure generation, and the preparation of the manuscript, with the following exceptions.

Markus Elze provided statistical consultation, aiding the refinement of an earlier model for footprint detection. Markus Elze also significantly contributed to the writing of Supplementary Sections 1.4, 1.6—10, and 1.13. Pierre Cauchy provided ongoing computational and analytical support and provided the rightmost panel of Figure S6 and the lower three panels of Figure S7. Peter N. Cockerill contributed the figure and figure legend for Figure S3 and along with Constanze Bonifer and Sascha Ott, provided ongoing general guidance. All authors provided comments and amendments to the manuscript during preparation and the peer review process.

Published online 25 September 2013

Nucleic Acids Research, 2013, Vol. 41, No. 21 ***e201***
doi:10.1093/nar/gkt850

Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data

Jason Piper[1,2], Markus C. Elze[1,3], Pierre Cauchy[2], Peter N. Cockerill[4,*], Constanze Bonifer[2,*] and Sascha Ott[1,*]

[1]Warwick Systems Biology Centre, University of Warwick, Coventry, CV4 7AL, United Kingdom, [2]School of Cancer Sciences, Institute of Biomedical Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, B15 2TT, United Kingdom, [3]Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom and [4]School of Immunity and Infection, Institute of Biomedical Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, B15 2TT, United Kingdom

Received May 7, 2013; Revised August 29, 2013; Accepted September 1, 2013

ABSTRACT

The expression of eukaryotic genes is regulated by *cis*-regulatory elements such as promoters and enhancers, which bind sequence-specific DNA-binding proteins. One of the great challenges in the gene regulation field is to characterise these elements. This involves the identification of transcription factor (TF) binding sites within regulatory elements that are occupied in a defined regulatory context. Digestion with DNase and the subsequent analysis of regions protected from cleavage (DNase footprinting) has for many years been used to identify specific binding sites occupied by TFs at individual *cis*-elements with high resolution. This methodology has recently been adapted for high-throughput sequencing (DNase-seq). In this study, we describe an imbalance in the DNA strand-specific alignment information of DNase-seq data surrounding protein–DNA interactions that allows accurate prediction of occupied TF binding sites. Our study introduces a novel algorithm, Wellington, which considers the imbalance in this strand-specific information to efficiently identify DNA footprints. This algorithm significantly enhances specificity by reducing the proportion of false positives and requires significantly fewer predictions than previously reported methods to recapitulate an equal amount of ChIP-seq data. We also provide an open-source software package, pyDNase, which implements the Wellington algorithm to interface with DNase-seq data and expedite analyses.

INTRODUCTION

The correct tissue-specific and temporal function of the genome is tightly controlled by transcription factors (TFs) that recognise specific DNA sequences and regulate the expression of specific genes. However, they do not act as single molecules but interact with each other to form large multi-protein assemblies that act as platforms for the recruitment of members of the epigenetic regulatory machinery (1,2). One of the significant challenges facing gene regulation studies is the identification of sites where TFs are bound to specific genes in a specific regulatory context. Although previous studies have shown a direct link between the sequence as well as tissue specificity of a number of TFs and gene expression patterns (3,4), the mechanisms behind how defined DNA sequences and the assembly of TF complexes translate into global gene expression patterns remains to be fully understood.

Characterising TF binding sites (TFBSs) across the entire genome is a monumental task. It is estimated that the total number of TFs in the human genome number ~1500, where several hundred of these may be active in a given cell type at any one time (5). Currently, the 'gold standard' for identifying occupied TFBSs in a given context uses chromatin immunoprecipitation paired with high-throughput sequencing (ChIP-seq) (6), which requires either a high-quality antibody or high cell numbers or alternatively epitope tagging. Although

*To whom correspondence should be addressed. Tel: +44 2476 150258; Email: s.ott@warwick.ac.uk
Correspondence may also be addressed to Constanze Bonifer. Tel: +44 121 414 8881; Email: c.bonifer@bham.ac.uk
Correspondence may also be addressed to Peter Cockerill. Tel: +44 121 414 6841; Email: p.n.cockerill@bham.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

ChIP-seq has proven to be extremely powerful, it is not without limitations: It is only possible to characterise one TF per experiment, it cannot be used alone to differentiate between primary and secondary binding (7), and the protein binding regions of the genome identified by ChIP-seq are in the order of several hundred base pairs. Progress has been made in this respect with the advent of ChIP-exo (8), which increases resolution of ChIP-seq data to below 50 bp, but this method has yet to see widespread adoption.

Another widely used approach in gene regulation studies uses DNase I as a tool to identify DNase I Hypersensitive Sites (DHSs) within chromatin (1). DHSs represent open chromatin regions that are normally only accessible at sites of active regulatory elements such as transcriptional enhancers. The recent development of DNase-seq has allowed more comprehensive mapping of the active chromatin landscape than is possible with ChIP-seq (9). The specific patterns of DNase I cleavage within DHSs also provide additional information about regions of DNA that are bound by proteins and are thereby protected from DNase I digestion, a feature that has been exploited for many years to obtain information about DNA–protein interactions at specific genes (10,11). However, the genome-wide data gained from this method are not trivial to analyse. DHSs can occupy hundreds of base pairs, and the entire complement of such sites contains an intrinsically high number of different specific TFBSs (9).

Although analyses of DNase-seq data were originally confined to identifying DHSs by peak detection, there have recently been several advances in the analysis of the raw tag counts that correspond to DNase activity at base pair resolution. The first of these digital genomic footprinting (DGF) methods were developed in yeast, where tag counts were processed with a rank transformation and tested for depletions in reads corresponding to occupied TFBSs using a binomial test (12). Subsequently, the first DGF studies in mammalian cells used a machine-learning approach where the tag counts were truncated, smoothed and differentiated, followed by the supervised training of a Hidden Markov Model on the known TFBSs in the *FMR1* promoter. Viterbi decoding was then performed to provide binary classifications (bound or unbound) for every base in the genome (13). Although several sets of footprints for various cell types as well as the model parameters were published, a software implementation was not made available. Another machine-learning approach, CENTIPEDE, trains an unsupervised Bayesian mixture model on the raw tag counts surrounding all genomic occurrences of a specified motif of interest to predict the binding states of each motif occurrence (14); however, unlike the previous methods, it cannot make predictions at arbitrary genomic loci. A software implementation of the CENTIPEDE algorithm is available but requires data to be pre-processed by the user into non-standard formats. The ENCODE project (15) has produced the most comprehensive set of DGFs in human cells by performing high-sequencing depth DNase-seq experiments on a multitude of cell types, adapting their previous footprinting methodology (12) to human data through the use of a metric that calculates the ratio of DNase-seq tags within a binding site to those directly outside (the Footprint Occupancy Score) (7).

Using publically available DNase-seq data from the ENCODE project, we describe how the alignment direction of DNA fragments relative to the reference strand exhibits a characteristic strand imbalance in the patterns surrounding known protein–DNA binding sites. We introduce Wellington, a novel footprinting algorithm that uses this knowledge to identify protein–DNA interactions in DNase-seq data with increased performance over previous methods, by reducing the number of false positives in our predictions. Alongside this, we provide the pyDNase software package to interface with DNase-seq data to run the Wellington algorithm and accelerate development of further analysis methods for these data. pyDNase and Wellington form a complete tool chain that can be used to identify protein–DNA interactions in any DNase-seq experiment performed according to the 'double-hit' protocol (16). Finally, we compared the performance of the different footprinting methods on a single data set, which we hope will be useful to the community in their decision of how to approach DGF tasks.

MATERIALS AND METHODS

Data

Aligned double-hit DNase-seq data and genomic co-ordinates of DHSs (K562: wgEncodeUwDgfK562, HepG2: wgEncodeUwDgfHepg2, A549: wgEncodeUwDgfA549, SkMC: wgEncodeUwDgfSkmcAln) and PhyloP conservation (Vertebrate phyloP46way) scores were downloaded from the UCSC genome browser (17). K562 data corresponding to the original single-hit DNase-seq library preparation method (9) were downloaded from the Sequence Read Archive (accession SRS131306) and aligned to hg19 using bowtie 1.0.0 (18) with the command line parameters '-a -best -strata -v 2 -m 1'. ChIP-seq data were downloaded as peaks from the ENCODE project's ChIP-seq studies (19); for track names, see Supplementary Table S1.

The Wellington algorithm

To detect protein–DNA binding sites, we must characterise the activity of DNase I and define what we consider to be a footprint. It is known that the activity of DNase I is lower in regions of inaccessible chromatin owing to protection of cleavage by histones or protein–DNA interactions. DNase I activity is therefore higher in regions of open chromatin without a bound protein. Protein–DNA binding sites can be detected by searching for a characteristic depletion of DNase I cuts compared with a large number of cuts in the surrounding region of open chromatin that do not harbour bound proteins.

To formalise our hypothesis test, we use the notation introduced in Figure 1. We will call the region surrounding the possible footprint the shoulder region. Let l_{FP} be the length (in base pairs) of the possible footprint and l_{SH} be the length (in base pairs) of the shoulder on each side of the possible footprint. We consider counts of cuts in these

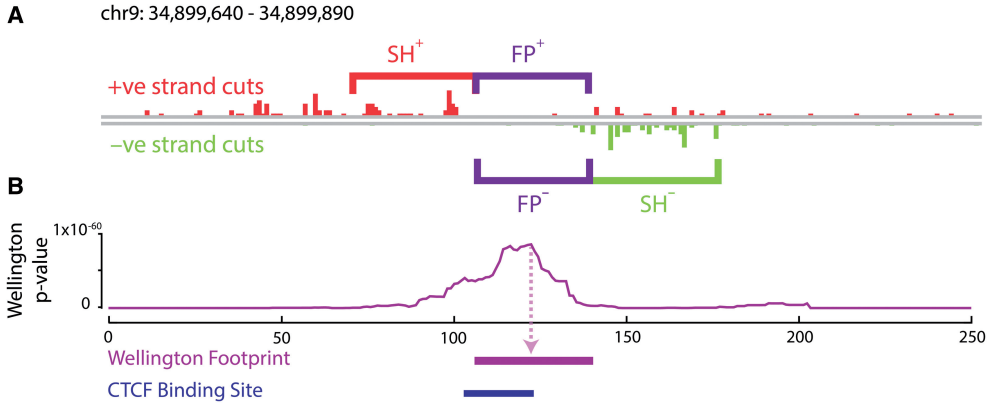


Figure 1. Wellington: a novel strand sensitive algorithm for the identification of protein–DNA binding sites from DNase-seq data. (**A**) The Wellington algorithm calculates *p*-values for every base pair in all DNase hypersensitive sites in a given DNase-seq data set, where the *s*-value is assigned to the base pair at the centre of the footprint. For each base pair, Wellington tests the hypothesis that there are significantly more reads aligning to the forward reference strand in the upstream shoulder region (SH^+) with respect to the +ve strand footprint region (FP^+) and significantly more reads aligning to the reverse reference strand in the downstream shoulder region (SH^-) with respect to the −ve strand footprint region (FP^-). (**B**) Example output of the Wellington algorithm. The corresponding footprint prediction recapitulates the ChIP-seq confirmed CTCF-binding site.

regions where 'cuts' refers to 5′ ends of the aligned sequencing tags. We consider four cut counts: the total number of cuts on the forward reference strand inside the possible footprint (FP^+), the cut count in the upstream shoulder region on the forward reference strand (SH^+), the cut count on the backward reference strand inside the possible footprint (FP^-) and the cut count in the downstream shoulder region on the backward reference strand (SH^-).

We now test the null hypothesis that the number of reads is proportional to the region length by using a binomial test. As the number of reads can depend on the strand, e.g. because the protein structure might be such that it only inhibits DNase I activity on one strand, we test both strands separately. We consider these tests to be independent, as each ~200 bp fragment will at most produce either one forward or one backward read close to the footprint site under investigation. With F(k, n, p) being the binomial cumulative distribution function, i.e. the probability of achieving at least k out of n successes with the probability of each success being p, we calculate a *p*-value using the formula *p*-value = $\{1\text{-}F[FP^+, FP^{++}SH^+, l_{FP}/(l_{FP}+l_{SH})]\} \cdot \{1\text{-}F[FP^-, FP^-+SH^-, l_{FP}/(l_{FP}+l_{SH})]\}$. This *p*-value is for a given possible footprint of length l_{FP} with surrounding shoulder regions of length l_{SH}.

We can calculate *p*-values for different possible footprint and shoulder lengths l_{FP} and l_{SH}. We can then choose which regions we wish to consider footprints by selecting an appropriate threshold for the *p*-values and subsequently using a greedy selection strategy for footprint identification. The parameters l_{FP} and l_{SH} are individually determined for each footprint using maximum likelihood estimation. The default values for l_{FP} are bound between 11 and 26 base pairs, whereas l_{SH} is fixed at 35 base pairs. Both of these parameters can be user-settable at run time with either ranges or fixed values. Further details are provided in the supplementary material.

Validation of predicted binding sites

We downloaded peaks determined by ENCODE's peak calling algorithm (specifically, ENCODE's 'optimal', high confidence set of peaks) for ChIP-seq experiments corresponding to a range of TFs. ChIP-seq confirmed binding sites were defined as motif instances falling within these peaks for each TF, and unbound motif locations were defined as motif instances falling outside ChIP-seq peaks.

To calculate ChIP-seq recapitulation, we used Wellington to calculate footprint *p*-values for each base pair in all DHSs and compared footprints with ChIP-seq positive motif instances. A ChIP-seq confirmed binding site is said to be successfully recapitulated by DNase-seq data if either at least 70% of the footprint is contained within the binding site or vice versa. This criterion is necessary as protection from DNase I is not always centred perfectly on a DNA motif. The same method was used when analysing Hesselberth *et al.* (12) footprints, Neph *et al.* (7) footprints and DHSs.

Average conservation scores were calculated using Vertebrate phyloP46way, and motif content was calculated using the genomic locations of 214 curated ChIP-seq verified position weight matrices published as part of the HOMER suite (20). For full details, see supplementary material.

RESULTS

Strand imbalance information increases the predictive power of footprinting algorithms

Strand-specific information in the context of DNase-seq data has been used primarily to describe TF-specific

cleavage patterns in reference to the orientation of a known DNA motif (13,14). Previous efforts at predicting DGFs have been strand-agnostic, ignoring alignment strand information and considering DNase I cleavage activity as absolute, without regard to the orientation of the sequenced fragment relative to the cut site. However, if one considers that the DNA fragments generated by DNase cutting are likely to originate predominantly from within DHSs, with a high probability of spanning occupied binding sites, then the strand to which the sequence tags align is likely to be highly informative with regard to the relative position of TFBSs. This is because the upstream end of a DHS fragment will be aligned as a +ve strand sequence tag, whereas the downstream end will be aligned as a −ve strand sequence tag, as illustrated in Supplementary Figure S3. Hence, for DNA fragments that span DHSs, and encompass DNase I footprints, the DNase I cuts identified from +ve strand alignments will be concentrated to the left, and those from −ve strand alignments will be concentrated to the right. Chromatin structure influences the digestion pattern, as there is a lower probability of sequencing DNA fragments that extend away from the DHS. This is caused by the fact that these fragments will be of lower abundance due to the lower probability of generating a second DNase I cleavage within flanking regions occupied by nucleosomes. Such fragments will thus likely be discarded during the necessary process of size selection before or during library preparation.

We tested the aforementioned predictions by considering the alignment strand when visualising DNase I cleavage sites in the vicinity of known motifs using published DNase-seq data at ChIP-seq verified binding sites from K562 cells that are available from ENCODE (Figure 2A and B). Similar to the imbalance of sequencing reads observed in ChIP-seq and DHS mapping (21), we noted that DNase-seq data surrounding binding sites often exhibit an abundance of sequencing reads aligning to the +ve reference strand upstream of the binding site, and reads aligning to the −ve reference strand downstream of the binding site, consistent with these tags representing opposite ends of DNA fragments spanning protected regions. This was particularly evident when DNase I cuts at binding motifs for specific factors across the genome were collapsed into a heat map (Figure 2B). When investigating a diverse set of TFs, we noticed that the imbalance varies in strength, with some binding sites having diminished strand imbalance, and others showing almost none. However, we never observe a 'reverse' imbalance of sequencing reads aligning to the −ve reference strand upstream of the binding site, and reads aligning to the +ve reference strand downstream of the binding site (Supplementary Figure S4). Although this imbalance is prominent in the data generated using the newer double-hit protocol used for all recent ENCODE DNase-seq data, the pattern is less pronounced in older data generated by the single-hit DNase-seq library preparation protocol (9) (Supplementary Figure S5).

It is also evident that more DNase I cut sites are detected immediately adjacent to the DNase I footprints, perhaps because the non-protected regions of a DHS are cleaved multiple times, with the smaller fragments being lost from the analysis. Overall, the number of reads aligning to the positive and negative strands in each DHS is roughly equal (Supplementary Figure S1) and so does not account for this imbalance. For some but not all motifs, additional information can be gained by re-orienting the DNase-seq data according to the orientation of the specific motif (Supplementary Figure S6). In the case of CTCF, a region of DNase I hypersensitivity exists on the −ve strand in a region that separates the major CTCF consensus motif from a secondary CTCF-binding site reported by others (13,22,23) When the motifs are aligned in the same orientation, this second site appears as a separate distinct protected region in Supplementary Figure S6. Here, we also show that CTCF motif scores are inversely correlated with Footprint Occupancy Scores, revealing that poorer motifs are less likely to generate clear footprints, as they are more susceptible to DNase I cleavage within the binding sites.

To assess whether the consideration of strand imbalance in DNase-seq data surrounding protein–DNA binding sites has an equally significant impact on the accuracy of DGF, we developed Wellington, a novel algorithm that performs DGF on DNase-seq data without the need for any prior knowledge, such as position weight matrices for the motifs that are likely to be annotated as a footprint. Wellington makes use of the sequence tag strand imbalance and searches DHSs for footprints that have a statistical enrichment of reads aligning to the +ve and −ve reference strand upstream and downstream of the binding site, respectively, with a depletion of reads on both strands in the region of the binding site. Figure 1 shows an example of such a footprint at a binding site for the TF CTCF containing a CTCF binding motif in the K562 data. This example demonstrates that Wellington footprints can accurately recapitulate the presence of a bound protein at a known TFBS.

To ensure that we were not missing genuine protein–DNA binding sites by excluding footprints that exhibited strand imbalance in the opposite direction, we again applied the Wellington algorithm to the ENCODE K562 DNase-seq data, but simultaneously applied it in a 'reverse' mode. This detected features exhibiting strand imbalance in the opposite direction to that which we demonstrated in Figure 2, (i.e. reads aligning to the −ve reference strand upstream of the binding site, and reads aligning to the +ve reference strand downstream of the binding site). Using the reverse Wellington method, we made footprinting predictions and compared them with those made by Wellington at the same *p*-value threshold of 1×10^{-30} (Figure 3A). All footprints identified possess the typical depletion in DNase I signal at the centre of the footprint (Figure 3B and D). As it is known that sequence conservation is correlated with the strength of TF binding (5,7,12–14), we investigated PhyloP (24) conservation scores surrounding footprints identified by both Wellington and reverse Wellington. We discovered that footprints only identified by Wellington showed an enrichment in sequence conservation at the centre of the footprint. This also held true for the footprints identified by both algorithms (due to there being sufficient reads on

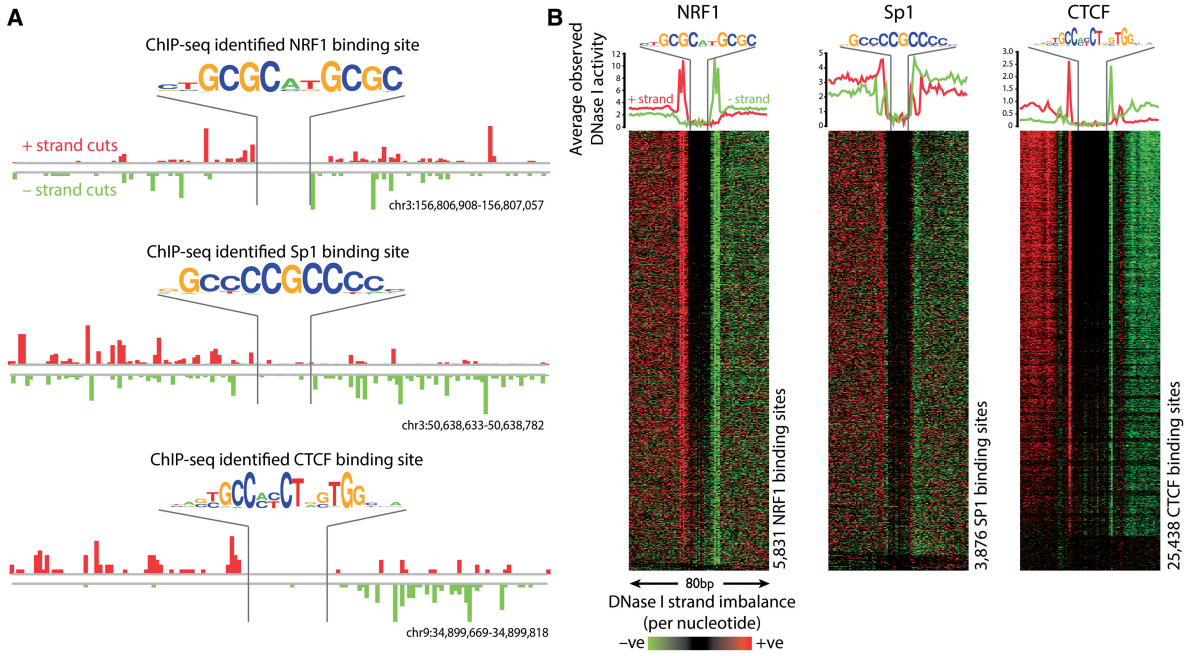


Figure 2. DNase I cleavage patterns surrounding known protein–DNA interactions as identified by ChIP-seq exhibit a strand imbalance, regardless of the strand where the binding motif is located. (**A**) Individual representative regions of DNase-seq data flanking NRF1, Sp1 and CTCF binding sites illustrate large numbers of sequencing fragments aligning to the positive reference strand upstream of the protein–DNA binding site and to the negative reference strand downstream of the protein–DNA binding site. These patterns exist independent of the direction in which the binding motif is located. (**B**) Heat maps show that the DNase-seq strand imbalance surrounding NRF1, Sp1 and CTCF binding sites identified by ChIP-seq exists on a genomic scale relative to the reference strand, irrespective of motif orientation (heat maps relative to motif orientation are shown in Supplementary Figure S4). Red indicates an excess of positive strand cuts over negative strand cuts per nucleotide position, and green indicates an excess of negative strand cuts. Binding sites are sorted from top to bottom in order of decreasing Footprint Occupancy Score (7).

both strands for both methods to detect a footprint). However, 'reverse footprints' identified by reverse Wellington only, did not show any evidence of enrichment in conservation score (Figure 3C), suggesting they are artefacts. To exclude the possibility that this result was only associated with the specific significance threshold chosen, we ran this analysis over a range of significance thresholds, but the main outcome of the analysis did not change (Figure 3E). Another indicator of the quality of footprint predictions, motif content (7,12–14), was also investigated. We found that motifs were enriched at the centre of footprint predictions (Supplementary Figure S7) and that over a range of significance thresholds, the pattern in the average motif content was the same as the average conservation score, with Wellington outperforming reverse Wellington (Figure 3F). Based on the fact that 'reverse footprints' with reverse strand imbalance patterns had very low motif content and very low average conservation scores, we consider these to be largely false positives. The majority of these are found adjacent to (5041, 54%), or in between (2734, 29%) footprints identified by Wellington (Supplementary Figure S8), with the minority (1607, 17%) having no neighbouring footprint within 50 bp. This indicates that these false positives are 'ghost' sites identified between or next to the shoulder regions of true footprints. To a strand-agnostic algorithm, these will appear to be depletions in DNase I activity associated with protein–DNA binding events. It is only by considering the strand information that it becomes possible to identify and discard them as artefacts in the data.

We next visualised footprints identified by Wellington at regions with known protein–DNA interactions that have previously been characterised by manual footprinting approaches, including the *FMR1* promoter (25), the IL-3 gene +4.9 kb CTCF site (26) and the β-globin LCR HS2 DHS (27). Figure 4 and Supplementary Figure S7 demonstrate the high precision with which Wellington infers regions of protein–DNA interaction.

Wellington is highly accurate at inferring protein–DNA interactions from DNase-seq data

To further assess the performance of the Wellington algorithm at identifying protein–DNA interactions compared with other methods, we used a range of different validation techniques, again using DNase-seq, footprinting and ChIP-seq data published by ENCODE. We also considered an implementation of Wellington that ignores strand information in the data, 'Wellington 1D' (see supplementary material for details), to assess the impact of the strand information on footprinting performance independently of the footprinting method. In the first instance, we compared our footprinting predictions for the K562

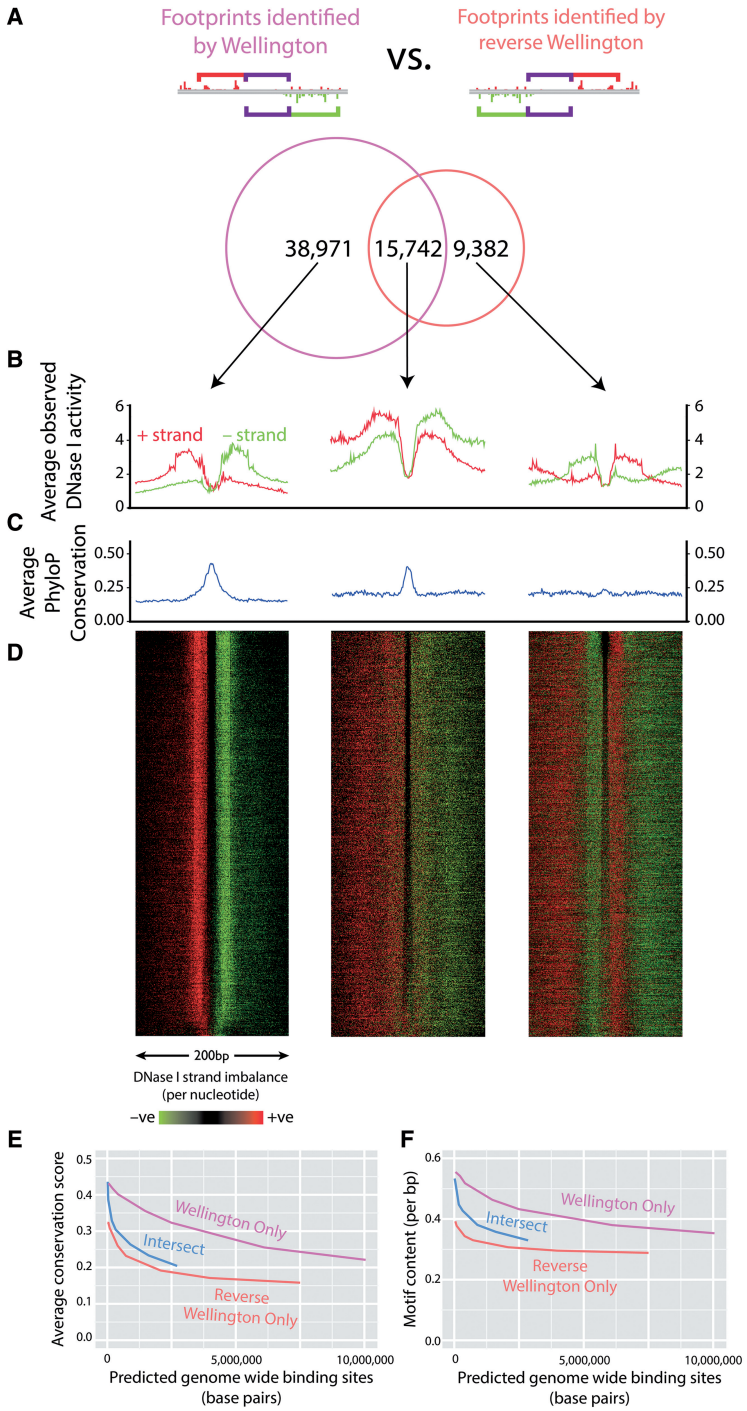


Figure 3. Strand imbalance information is crucial for the identification of true protein–DNA interactions. (**A**) The Wellington algorithm was run on K562 DNase-seq data in parallel with a modified version of the Wellington algorithm (reverse Wellington) designed to identify strand imbalance in the opposite direction than expected, i.e. reads aligning to the negative reference strand upstream of the binding site, and reads aligning to the positive reference strand downstream of the binding site. (**B**, **C**) Although footprints identified only by reverse Wellington harbour the characteristic depletion of DNase I cleavage, we find that they do not exhibit the increase of conservation typical of known protein–DNA interactions (7,12–14). (**D**) Heat maps of the DNase I signal surrounding the reverse Wellington footprints support the hypothesis that false-positive footprint signals primarily arise from junctions in between adjacent protein–DNA binding sites. (**E**) The observation of low conservation scores of footprints detected by reverse Wellington is maintained when comparing Wellington and reverse Wellington footprints at a range of significance levels. (**F**) Footprints detected by reverse Wellington contain fewer TF-binding motifs.

Figure 4. Wellington footprints recapitulate known protein–DNA interactions at (**A**) the *FMR1* promoter (25), (**B**) the *IL3* +4.9 kb insulator (26) and (**C**) the β-globin HS2 hypersensitive site (27) and refine previous footprinting predictions at these loci (7).

DNase-seq data with K562 ChIP-seq data for a range of TFs (ATF3, c-Myc, CTCF, JunD, Max, NFE2, NRF1, NRSF, PU.1, Sp1 and USF1). We investigated the ChIP-seq recapitulation performance of our method by searching for motifs within footprints using a range of decreasing stringencies for the footprint *p*-value (Figure 5A). Over all stringencies, Wellington performed the best, meaning that the efficiency of Wellington at recapitulating ChIP-seq data per base pair of prediction was higher than that of other methods. For example, it required approximately 60% fewer predictions compared with Neph *et al.*'s footprint analysis to recapitulate an equal amount of ChIP-seq data for these 11 TFs. Although this analysis clearly showed the increased coverage gained by Wellington, it did not take the number of false positives or false negatives made by these predictions into account. To address this, we calculated the Average Nucleotide Performance Coefficients (28) for the 11 ChIP-seq experiments as a function of total genomic footprint predictions, which revealed a consistently higher correlation between the ChIP-seq confirmed binding sites and the Wellington footprints across all sensitivities compared with other methods (Figure 5B).

A validation method commonly used in classification experiments, the Receiver Operator Characteristic (ROC), assesses the performance of a binary classifier over a range of significance thresholds (see supplemental material for details). Wellington yielded an area under the ROC curve higher than 0.80 in the ability to recapitulate all 11 TFs in K562 cells (Figure 5C), indicating that Wellington is an excellent predictor of TF binding (29). ROC analysis was also performed on HepG2 and A549 DNase-seq data (Supplementary Figures S10 and S11), yielding similar performance. Although this method has been used in the validation of previous footprinting methods, it should be noted that due to the relatively small number of true positives (bound motif instances) and large number of true negatives (unbound motif instances) in the genome for most TFs (Supplementary Table S1), this statistic is skewed towards assessing the ability of an algorithm to correctly predict unbound locations.

CENTIPEDE (14) is based on known binding motif locations and learns one footprint model for each individual motif. It is therefore capable of using features of footprints that are specific to one or few motifs. In contrast, Wellington is a generic footprinting method for the detection of a wide range of binding sites. It does not depend on previous knowledge of motifs and does not learn models for individual motifs. We therefore considered the possibility that CENTIPEDE might outperform Wellington. However, we found that Wellington still outperformed CENTIPEDE when comparing the Positive Predictive

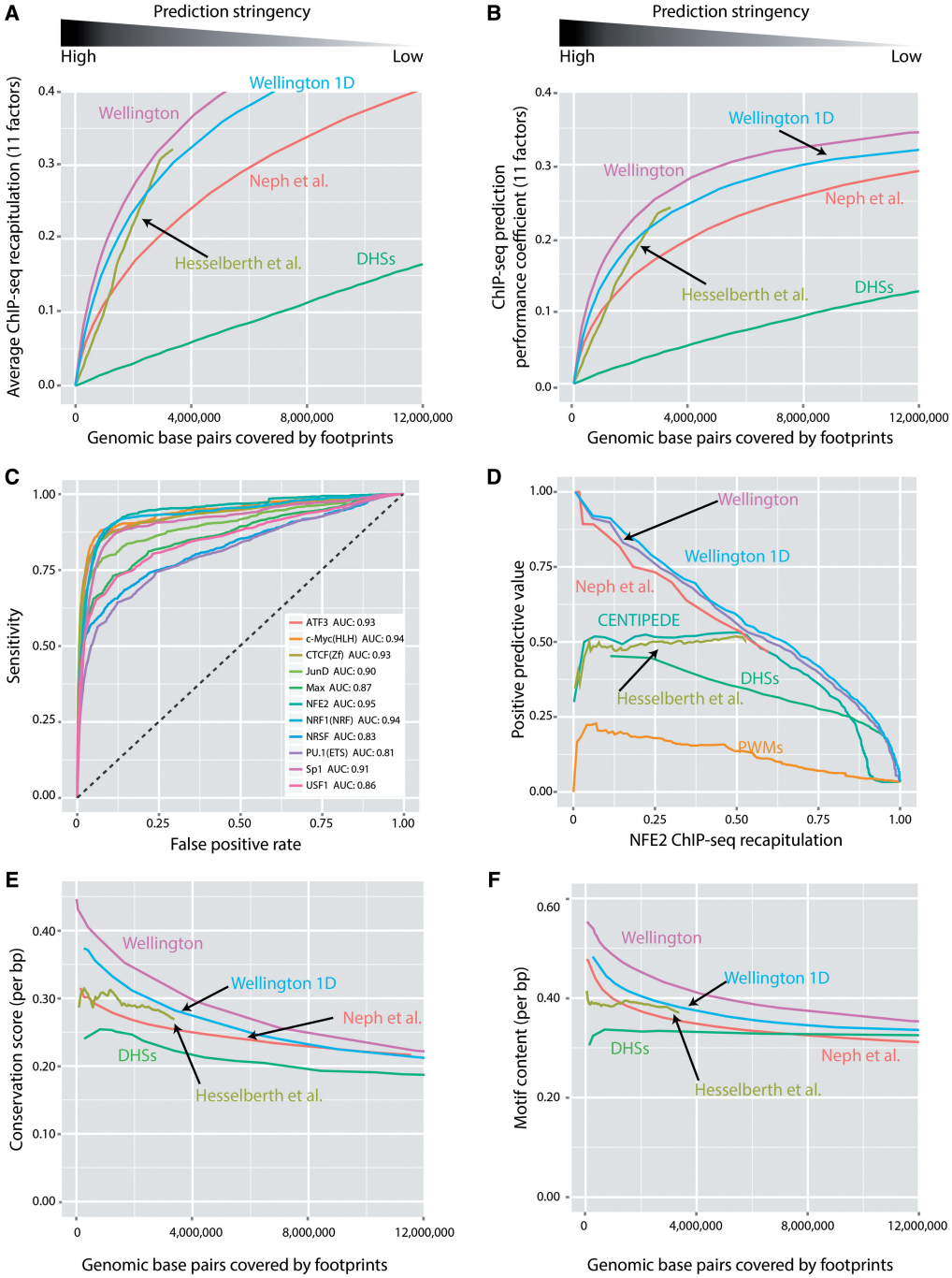


Figure 5. Wellington outperforms other methods with respect to ChIP-seq recapitulation performance, sequence conservation and motif content within footprints. (**A**) Wellington is able to recapitulate a given amount of ChIP-seq data with approximately half the number of genomic predictions compared with Neph *et al.* (7). The horizontal axis shows the total number of base pairs in the genome that are covered by footprints at a given stringency, the vertical axis shows the average performance of these footprints in recapitulating binding sites found from ChIP-seq data for 11 TFs in K562 cells. DHSs: using DNase hypersensitive sites to recapitulate ChIP-seq binding sites. (**B**) The nucleotide performance coefficients for these predictions (28) take numbers of false positives and false negatives into account and show a consistent finding compared to (A). (**C**) ROC curves for Wellington binding site predictions of 11 genomic TFs. The dashed line shows the expected performance of a random classifier. AUC: Area under curve. (**D**) Using the NFE2 ChIP-seq data as an example, we illustrate that the positive predictive value (the proportion of binding site predictions that are correct) of Wellington is either equal to or exceeding other footprinting techniques. (**E**, **F**) Wellington footprints have consistently the highest PhyloP conservation scores and motif content.

Value (the fraction of predicted binding sites that are confirmed in ChIP-seq data, PPV) as a function of the ChIP-seq coverage (Figure 5D), implying that Wellington can be specifically used for the purpose of determining *in vivo* occupancy of a given motif. The method by Neph *et al.* and Wellington performed comparably when the location of a binding motif was known, but CENTIPEDE's Positive Predictive Value was lower at lower sensitivities. Comparable results were observed for the other 10 TFs (Supplementary Figure S9). However, it is worth noting that when performing analyses that require the presence of a motif in the footprint, a high number of motif-less footprints are masked and unknown motifs are not found. Moreover, the assumption that a given TF generates a uniform digestion pattern limits the predictive power of the algorithm, for example, it has been shown that multiple clusters of DNase I cleavage patterns exist for CTCF (13). In addition, the dynamic binding behaviour of a specific TF can be modulated by interaction with other factors binding within the DHS (30). The extent of this has not yet been investigated, and other TFs could also generate differing DNase I cleavage patterns dependent on differing binding dynamics at individual sites across the genome.

All of the aforementioned analyses rely on ChIP-seq data as a gold standard, and therefore false positives in ChIP-seq analyses can appear as false negatives in footprinting assays and *vice versa*. Other metrics that do not rely on ChIP-seq data, such as conservation scores and motif enrichment, which are also highly correlated with TF binding and regulatory activity (31), can be used to assess footprinting performance. We therefore calculated the average PhyloP conservation score and the average motif content of footprints across a range of thresholds on footprint *p*-values. To calculate motif content, we used a library of 214 ChIP-seq derived DNA motifs. Across all sensitivities, Wellington footprints yielded higher conservation scores and motif content per base pair (Figure 5E and F) than other methods, further demonstrating Wellington's ability to identify footprints enriched for regulatory elements with high conservation scores and protein binding potential. This notion is exemplified in Supplementary Figure S12, which depicts the DHS at the *FMR1* promoter demonstrating the precise overlap of regions with high footprinting *p*-values and high conservation scores. The ability for Wellington to outperform Wellington 1D in these metrics confirms that the consideration of the strand information in DNase-seq data assists in reducing the number of low conservation scoring false-positive 'reverse' footprints in the genome without affecting predictive power. When considering data generated with the original single-hit protocol, however, we found that Wellington did not improve over Wellington 1D (Supplementary Figures S14–S16). This is likely due to the fact that the single-hit data have less pronounced strand imbalance patterns (Supplementary Figure S5), which Wellington is specifically designed to detect.

In summary, Wellington efficiently increases the specificity of footprint detection by avoiding artefacts, which only become apparent when considering the alignment strand of DNase I cuts in DNase-seq data (Supplementary Figure S13). It therefore maintains excellent ChIP-seq recapitulation performance whilst significantly reducing the total number of predicted footprints in the genome.

pyDNase: a Python package for analysing DNase-seq data

At present, no free open source software package is available that would allow the analysis of DNase-seq data with the aim of performing digital footprinting without specifying any prior parameters, such as motif of interest. DGF presents unique challenges in data handling due to the large (>500 million) number of reads, and the necessity to interact directly with raw alignment data to perform complex analyses. With ChIP-seq, this step is unnecessary after basic peak calling and generation of extended read densities. We therefore developed pyDNase as the first open source DNase-seq analysis software package. pyDNase complements other common bioinformatics tools to establish the first functional DNase-seq footprinting pipeline. It is written in Python for higher-level functions and C for lower-level performance-critical functions. The analysis pipeline using pyDNase is outlined in Figure 6, whereby pyDNase serves a conduit between the raw alignment data and DNase-seq analysis algorithms such as Wellington. The most basic usage, a footprinting analysis with the default parameters can be performed by running the `wellington_footprints.py` script with the sequencing reads in BAM format, a list of DHSs in the data set, and an output location for the results (e.g. `$ python Footprint.py reads.bam dhs.bed ~/results/`), which will then output the footprint scores as a wig file, and footprints at various *p*-value cutoffs. The behaviour of this script is highly configurable through command line arguments. pyDNase allows Wellington footprinting of all DHSs in a 600 million read DNase-seq experiment in ~4–10 h on a desktop computer with 1 Gb of RAM and a 2.3 GHz Intel Core i5 processor. This will simplify and expedite data analyses as well as method development for future studies. pyDNase and the Wellington algorithm are available as a Python package, along with sample data sets, a step-by-step tutorial, and documentation of every method and class at http://jpiper.github.com/pyDNase and is freely released under the GNU GPLv3 open source software license.

DISCUSSION

By designing the Wellington algorithm to identify footprints using the knowledge that strand imbalance surrounds known protein–DNA interactions, we have increased our ability to perform DGF by reducing the number of motif-depleted non-conserved false positives. Footprints identified by Wellington show consistently higher average conservation scores, motif content and ChIP-seq recapitulation per base pair than other methods. Considering that the ChIP-seq recapitulation performance was the justification behind the previous claim of 0.4 to 2.3 million genomic footprints (dependent on the cell type) (7),

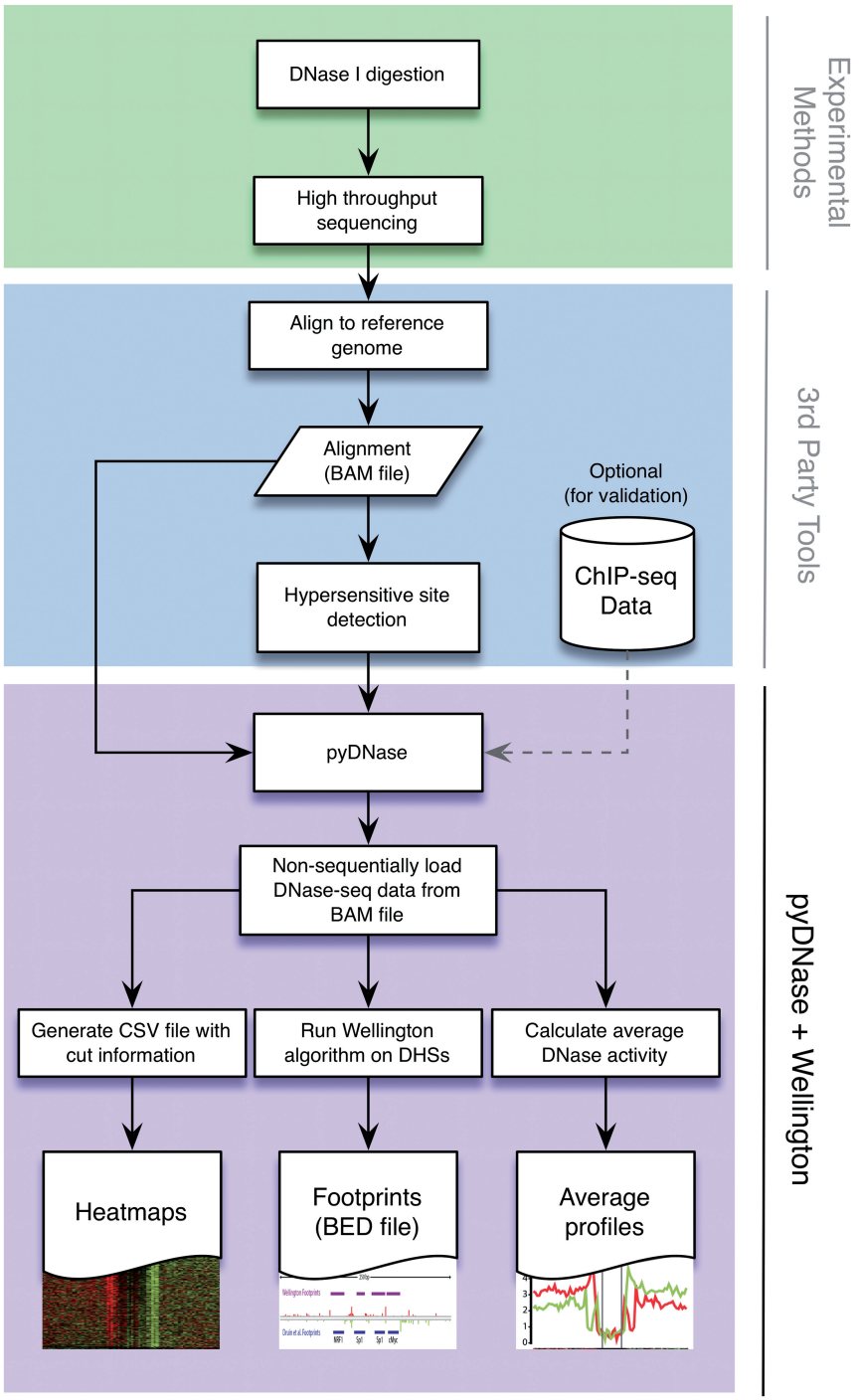


Figure 6. The pyDNase Python package forms a complete toolchain for the rapid analysis and footprinting of DNase-seq data. Using mapped DNase-seq reads as a BAM file, pyDNase not only has scripts to perform common analyses (heat maps, footprinting, average profiles) but also exposes an API to allow the easy development of further DNase-seq analysis tools.

the results presented here suggest that much less of the genome may be involved in protein-binding events than previously predicted. Wellington required approximately 60% fewer predictions compared with Neph *et al.*'s footprint analysis to recapitulate an equal amount of ChIP-seq data for 11 TFs. This is due to the large number of motif-less false positives in the Neph *et al.* set of predictions that do not impact on the chosen validation metrics. However, it remains difficult to determine exactly how many binding sites there may actually be as human DGF is still limited by sequencing depth (7) (Supplementary Figure S2).

We hypothesise that the strand imbalance is a natural consequence of the size selection step of the 'double-hit' protocol, which purifies ~50–200 base pair DNA fragments produced by DNase I digestion (Supplementary Figure S3). This is strengthened by the result that consideration of strand information does not contribute any predictive power to data generated by the single-hit DNase-seq method, which does not use size selection in the library preparation (9) and has detectable but less pronounced strand imbalance patterns (Supplementary Figures S5, S14–S16). In the double-hit protocol, eliminating the smallest digestion products and excluding larger chromatin fragments creates a bias towards sequencing DNA fragments that actually span the DNase I footprints where TFs are bound. Because the +ve and −ve strand sequence tags simply represent the opposite ends of the same sets of DNA fragments, this is a straightforward predictor of the location of a footprint relative to the 5′ end of the sequence tag. Giving due consideration to the introduction of strand imbalance surrounding sites protected by protein–DNA interactions in the double-hit DNase-seq data allows the development of analyses that reduce the number of false positives in footprint predictions.

This increased footprinting precision as well as the ability of Wellington to be used on *a priori* defined motifs opens the door to higher-order analyses, such as *de novo* identification of occupied *cis*-regulatory modules, as well as the elucidation of direct or indirect TF interaction in a given complex *via* determination of specific motif distances. Furthermore, the strand-specific cleavage patterns surrounding motifs bound by different TF families seemingly constitute unique, individual signatures, which may permit motif identification based solely on DNase-seq data.

The identification of TFBSs bound in a cell-type and cell-stage specific fashion is a key stage in gaining an understanding of differential gene expression underlying all cell differentiation processes. Using techniques such as DNase-seq, ChIP-seq, and algorithms such as Wellington, we can begin to document the TF-binding events that confer cell identity, developmental processes or which underpin aberrant regulation in diseases such as cancer. By significantly reducing the number of false-positive predictions, we decrease the need for multiple technical and biological replicates, which can be difficult to obtain for primary tissues such as patient samples. This opens up the possibility of performing analyses on disease-specific transcription regulation mechanisms, which have previously only been possible using data combined from multiple experiments over large numbers of cell lines (7,13).

It remains to be seen how footprinting algorithms can be further enhanced. Even though it is known that the pattern of the DNase-seq signal surrounding protein–DNA binding events is TF dependent, we found Wellington to perform well using a single model to search for all possible TF-binding events in a DNase-seq data set. The use of more complex mixture models could yield even better performance, which at some stage may even allow the incorporation of an analysis of the chromatin landscape. The speed at which new computational analyses of DNase-seq data are being developed is greatly surpassed by the rate at which new DNase-seq data are being generated (32). To encourage further investigations, we have released pyDNase and Wellington as a Python package for the fast and easy analysis of DNase-seq data. We hope that accelerates both the analysis of DNase-seq data and the development of advanced footprinting algorithms.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online, including [33].

ACKNOWLEDGEMENTS

The authors would like to thank Jessica 'Maverick' Hearn and the anonymous reviewers for their constructive feedback on the manuscript.

FUNDING

Engineering and Physical Sciences Research Council [EP/P50578X/1 PhD grant to J.P.] (in part), a Chancellor's Scholarship from the University of Warwick and a PhD Fellowship from the German National Academic Foundation (to M.C.E); Leukaemia & Lymphoma Research (to C.B. and P.N.C.) as well as from the Biotechnology and Biological Sciences Research Council [BB/I001220/1 to C.B.]. Funding for open access charge: RCUK block funding to University of Warwick Library.

Conflict of interest statement. None declared.

REFERENCES

1. Cockerill,P.N. (2011) Structure and function of active chromatin and DNase I hypersensitive sites. *FEBS J.*, **278**, 2182–2210.
2. Spitz,F. and Furlong,E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–626.
3. Zinzen,R.P., Girardot,C., Gagneur,J., Braun,M. and Furlong,E.E.M. (2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, **462**, 65–70.
4. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
5. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.

6. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
7. Neph,S., Vierstra,J., Stergachis,A.B., Reynolds,A.P., Haugen,E., Vernot,B., Thurman,R.E., John,S., Sandstrom,R., Johnson,A.K. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.
8. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.
9. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
10. Galas,D.J. and Schmitz,A. (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.*, **5**, 3157–3170.
11. Tagoh,H., Cockerill,P.N. and Bonifer,C. (2006) *In vivo* genomic footprinting using LM-PCR methods. *Methods Mol. Biol.*, **325**, 285–314.
12. Hesselberth,J.R., Chen,X., Zhang,Z., Sabo,P.J., Sandstrom,R., Reynolds,A.P., Thurman,R.E., Neph,S., Kuehn,M.S., Noble,W.S. *et al.* (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.
13. Boyle,A.P., Song,L., Lee,B.K., London,D., Keefe,D., Birney,E., Iyer,V.R., Crawford,G.E. and Furey,T.S. (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.*, **21**, 456–464.
14. Pique-Regi,R., Degner,J.F., Pai,A.A., Gaffney,D.J., Gilad,Y. and Pritchard,J.K. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.
15. ENCODE Project Consortium. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
16. Sabo,P.J., Kuehn,M.S., Thurman,R., Johnson,B.E., Johnson,E.M., Cao,H., Yu,M., Rosenzweig,E., Goldy,J., Haydock,A. *et al.* (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat. Methods*, **3**, 511–518.
17. Rosenbloom,K.R., Sloan,C.A., Malladi,V.S., Dreszer,T.R., Learned,K., Kirkup,V.M., Wong,M.C., Maddren,M., Fang,R., Heitner,S.G. *et al.* (2012) ENCODE data in the UCSC genome browser: year 5 update. *Nucleic Acids Res.*, **41**, D56–D63.
18. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
19. ENCODE Project Consortium, Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
20. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
21. Pepke,S., Wold,B. and Mortazavi,A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, **6**, S22–S32.
22. Bowers,S.R., Mirabella,F., Calero-Nieto,F.J., Valeaux,S., Hadjur,S., Baxter,E.W., Merkenschlager,M. and Cockerill,P.N. (2009) A conserved insulator that recruits CTCF and cohesin exists between the closely related but divergently regulated interleukin-3 and granulocyte-macrophage colony-stimulating factor genes. *Mol. Cell. Biol.*, **29**, 1682–1693.
23. Schmidt,D., Schwalie,P.C., Wilson,M.D., Ballester,B., Gonçalves,A., Kutter,C., Brown,G.D., Marshall,A., Flicek,P. and Odom,D.T. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, **148**, 335–348.
24. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
25. Drouin,R., Angers,M., Dallaire,N., Rose,T.M., Khandjian,E.W. and Rousseau,F. (1997) Structural and functional characterization of the human FMR1 promoter reveals similarities with the hnRNP-A2 promoter region. *Hum. Mol. Genet.*, **6**, 2051–2060.
26. Bowers,S.R., Calero-Nieto,F.J., Valeaux,S., Fernandez-Fuentes,N. and Cockerill,P.N. (2010) Runx1 binds as a dimeric complex to overlapping Runx1 sites within a palindromic element in the human GM-CSF enhancer. *Nucleic Acids Res.*, **38**, 6124–6134.
27. Elnitski,L., Miller,W. and Hardison,R. (1997) Conserved E boxes function as part of the enhancer in hypersensitive site 2 of the beta-globin locus control region. Role of basic helix-loop-helix proteins. *J. Biol. Chem.*, **272**, 369–378.
28. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
29. David,W., Hosmer,J., Lemeshow,S. and Sturdivant,R.X. (2013) *Applied Logistic Regression*. Wiley, Hoboken, New Jersey, USA.
30. Voss,T.C., Schiltz,R.L., Sung,M.H., Yen,P.M., Stamatoyannopoulos,J.A., Biddie,S.C., Johnson,T.A., Miranda,T.B., John,S. and Hager,G.L. (2011) Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. *Cell*, **146**, 544–554.
31. Xie,X., Mikkelsen,T.S., Gnirke,A., Lindblad-Toh,K., Kellis,M. and Lander,E.S. (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl Acad. Sci. USA*, **104**, 7145–7150.
32. Madrigal,P. and Krajewski,P. (2012) Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. *Front. Genet.*, **3**, 230.
33. Koohy,H., Down,T.A. and Hubbard,T.J. (2013) Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme. *PLoS One*, **8**, e69853.

Supplementary Methods, Discussion, and Figures for *Wellington: A novel method for the accurate identification of digital genomic footprints from DNase-seq data*

Jason Piper, Markus C. Elze, Pierre Cauchy, Peter N. Cockerill, Constanze Bonifer, and Sascha Ott

28th August 2013

1 Development of the Wellington Algorithm

1.1 Goals and underlying assumptions

The goal of Wellington is to statistically identify footprints from DNase I cut data. We define a footprint as a region where the number of DNase I cuts per base pair is significantly lower than in the surrounding area. Specifically, we compare the number of DNase I cuts per base pair inside the possible footprint region on the forward reference strand with cuts per base pair the upstream region on the forward reference strand and the number of DNase I cuts per base pair inside the possible footprint region on the backward reference strand with cuts per base pair the downstream region on the backward reference strand.

As stated in the Methods, we assume that the number of DNase I cuts is much lower (depleted) in regions of closed chromatin or of open chromatin with a bound protein than in regions of open chromatin without a bound protein, which is well established by the literature (Sabo et al., 2006; Boyle et al., 2008; Hesselberth et al., 2009; Cockerill, 2011; Neph et al., 2012). Thus, protein-DNA binding sites can be detected by finding a characteristic depletion of DNase I cuts compared to the surrounding region of open chromatin without bound proteins. Furthermore, we assume that the number of DNase I cuts in open chromatin without bound proteins is roughly proportional to the length of the

region. Thus, we can test if a region has a significantly lower than expected number of cuts to identify footprints or putative protein-DNA binding sites.

1.2 Rationale

It has been previously established that DNase I cuts are not distributed uniformly across the open chromatin, but that the probability is dependent on the DNA sequence around and at the cleavage site. This problem could theoretically be compensated by appropriate pre-processing, such as adjusting the observed number of reads by the DNase I cutting rates for the surrounding base doublet (Koohy et al., 2013). However, in practice, we and others have found that differences in cutting preferences are sufficiently small to not have an undue impact on the footprint identification. The difference in DNase I accessibility within and outside DHS in chromatin by far exceeds sequence dependent differences in the digestion frequency of naked DNA (Neph et al., 2012; Hesselberth et al., 2009).

Due to the methodology employed, DNase-seq typically only includes DNA fragments of a certain size range. Any fragments smaller than about 50bp or larger than about 250bp are discarded. Thus, DNase I needs to cut the DNA twice in reasonably close proximity for the fragment to be included in the analysis. This means that regions with multiple bound proteins in close proximity to each other or regions with bound proteins close to nucleosomal chromatin might be hard to identify. In order to avoid this problem as much as possible, we only consider cuts arising from fragments that span across the footprint site, i.e. those upstream of the footprint site on the forward strand and those downstream of the possible footprint site on the reverse strand. In principle, other bound proteins very close to the possible footprint site might still be a problem even with this step. However, this was not observed in practice.

The DNA fragment may be further shortened by additional DNase I cuts. This means that we typically expect to observe more shorter DNA fragments than longer DNA fragments. This can lead to an increased number of cuts just outside of footprints upstream on the forward strand and downstream on the reverse strand. Wellington currently does not explicitly utilise this phenomenon, as we did not observe it for all footprints.

1.3 Data preprocessing

Different DNase-seq techniques can produce sequencing artefacts, e.g. in the form of read spikes at single base pairs. If possible, appropriate preprocessing should be used in order to reduce the impact of sequencing artefacts and other undesirable phenomena mentioned above. Which preprocessing method is appropriate depends on the precise experimental method. Some authors have suggested to allow a maximum number of reads per base pair to reduce the impact of spikes. Whilst the provided software implementation of the Wellington algorithm offers this feature, we do not generally recommend it, as a lot of data may be needlessly discarded.. Rather, we encourage researchers either to identify and remove sequencing artefacts manually or to choose a sufficiently high p-value cutoff and to check for problems by shuffling the data and searching for footprints again (see below). Usually, the forward and the reverse strands will have a similar number of cuts (Figure S1). Thus, finding many more cuts on one strand than on the other in a region can be an indicator that a sequencing artefact may be present.

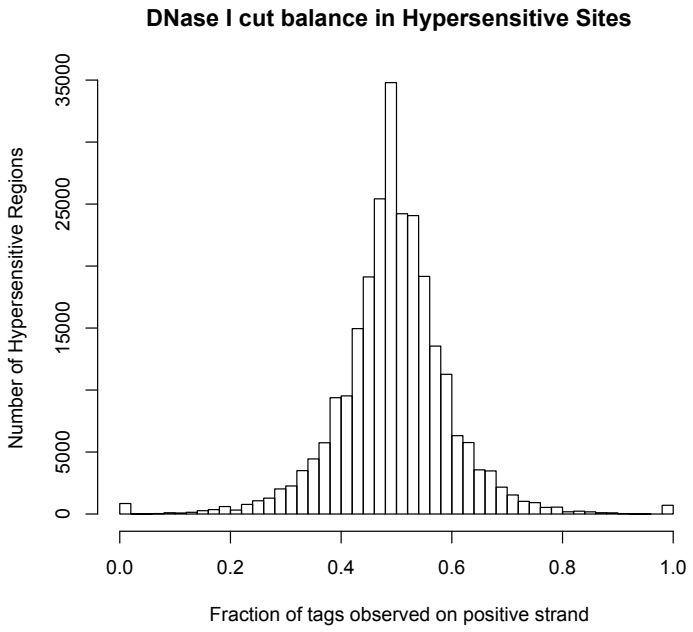


Figure S1: The numbers of cuts observed on the positive and on the negative strand are of the same order of magnitude. Within each hypersensitive region in K562 cells, the ratio of cuts on the positive strand to the total number of cuts has a mean of 0.50 and a standard deviation of 0.10.

1.4 Calculating a p-value for a given potential footprint

As described in the Methods, we use the following notation introduced in Figure 1. We call the region surrounding the possible footprint the shoulder region. Let l_{FP} be the length of the possible footprint and l_{SH} be the length of the shoulder on each side of the possible footprint. For now, consider l_{FP}, l_{SH}, and the centre of the potential footprint as given. We can then calculate the four DNase I cut counts that are relevant for the hypothesis test: the total number (i.e. sum over all base pairs) of cuts on the forward reference strand inside the possible footprint (FP^+), the number of cuts in the upstream shoulder region on the forward reference strand (SH^+), the number of cuts on the backward reference strand inside the possible footprint (FP^-), and the number of cuts in the downstream shoulder region on the backward reference strand (SH^-).

We test the null hypothesis that the number of cuts is proportional to the region length by using a binomial test. Because the number of cuts might depend on the strand, e.g. because the protein structure might be such that it only inhibits DNase I activity on one strand, we test both strands separately. With $F(k, n, p)$ being the binomial cumulative distribution function (the probability of achieving at least k out of n successes for the probability of each success being p), we calculate a p-value using the formula

$$p{-}value = F(FP^+, FP^+{+}SH^+, \frac{l_{FP}}{l_{FP}+l_{SH}}){*}F(FP^-, FP^-{+}SH^-, \frac{l_{FP}}{l_{FP}+l_{SH}})$$

This p-value is for a given possible footprint of size l_{FP} with surrounding shoulder regions of size l_{SH}.

1.5 A strand agnostic Wellington

In order to investigate the impact of the strand information of Footprinting results independently of footprinting methodology, we utilised a simplified version of Wellington which uses data on both strands, Wellington 1D. We calculate parameters in the model differently than above to account for this. Let the total number of cuts on both strands inside the possible footprint (FP), the number of cuts in the upstream shoulder region on the both strands (SH^u), and the number of cuts in the downstream shoulder region on both strands (SH^d). We then calculate a p-value using the formula

$$p - value = F(FP, FP + SH^d + SH^u, \frac{l_{FP}}{l_{FP} + l_{SH}})$$

1.6 Selecting footprint and shoulder widths at a given position

For a given centre position of a possible footprint, we can vary both the length of the possible footprint l_{FP} and the length of the shoulder l_{SH}. This results in a multitude of hypothesis tests, all of which may have different p-values. Typically, the researcher will specify a range of possible values for l_{FP} and l_{SH} that are appropriate. If no hypothesis tests are significant at the chosen significance threshold, there is clearly no evidence for this site being a footprint. If only one hypothesis test is significant at the chosen significance threshold, we consider this a footprint of length l_{FP} belonging to the significant test. Matters get slightly more complicated if more than one test is significant.

If more than one test is significant, we have successfully rejected multiple slightly different hypotheses. For practical purposes, we wish to have a set footprint length l_{FP} and shoulder length l_{SH} instead of multiple possible values. To achieve this, we choose the l_{FP} and l_{SH} that provide the most evidence against the null hypothesis and result in the lowest p-value. From a Bayesian perspective, this corresponds to putting a uniform prior over the previously specified ranges of possible l_{FP} and l_{SH}. It is straightforward to extend Wellington to allow arbitrary priors for l_{FP} and l_{SH}.

1.7 Greedy selection of footprints from all possible candidates

The previous section produced one p-value for each possible footprint centre base pair along with corresponding footprint widths. If we only have one significant p-value in a region, the corresponding possible footprint will be considered our one true footprint. Multiple significant p-values in a region may result in overlapping footprints and a decision has to be made how to deal with this phenomenon.

While overlapping footprints can occur and we wish to allow this, we also want to avoid artificially extending footprints simply for the reason that base pairs slightly away from the centre of a protein binding site will often still succeed in rejecting the null hypothesis. Thus, we wish to require two overlapping

footprints to overlap for less than a certain user-settable percentage, which defaults to 50%. This requirement in no way restricts what footprint patterns are possible. Lifting this requirement results in noticeably larger footprints, which are often somewhat longer than the desired maximum length for a single footprint.

To achieve this goal, we implement a greedy selection strategy. For a given region, we start by choosing the footprint with the lowest p-value as our first footprint, as this footprint offers the strongest evidence against the null hypothesis. After we have added this footprint to our list of identified footprints, we then consider any base pairs contained in this footprint not to be eligible to be the centre of additional footprints. For the next footprint, we continue in the same fashion by choosing the footprint with the lowest p-value, adding it to our list, and removing all base pairs contained in it from the list of possible footprint centres. This process continues until no eligible base pairs remain with a p-value below the significance threshold.

1.8 Choosing a significance threshold and assessing possible false positives

To choose a significance threshold, the fact that possibly billions of hypothesis tests are performed needs to be considered. We decide to err on the side of caution and perform a Bonferroni correction. To make the multiple testing correction as simple as possible for the end user, we adjust all p-values instead of just adjusting the significance threshold internally. More advanced methods, such as a Bonferroni-Holm correction, are not used for the sake of computational simplicity.

As, even with excellent preprocessing, the cut counts in open chromatin regions without bound proteins will be neither uniformly nor independently distributed, we typically recommend being more conservative than the standard $p < 0.05$ threshold (corresponding to 1.3 on the $-\log$ scale). For ENCODE datasets, we found that thresholds of $1.3 - 20$ work very well, depending on the desired number of false positives.

Ultimately, when applying this method to a dataset, we wish to adjust the p-value threshold we choose for calling footprints on a hypersensitive site-wise basis to generate a single set of footprints for the dataset (and not set a single p-value cutoff for the entire dataset). In order to do this, Wellington has a command line argument to employ an empirical method of estimating

the False Discovery Rate (FDR) as described previously (Neph et al., 2012; Hesselberth et al., 2009). Briefly, we shuffle the number of tags aligned to each base pair within a hypersensitive site and recalculate the footprint scores on this shuffled data 500 times, and then can determine a p-value threshold which would only occur at most 1 in 100 times, corresponding to an FDR of 0.01.

1.9 Sequencing Depth

The number of footprints called will largely be determined by the sequencing depth of the dataset. A common question from experimentalists remains 'How deep do I need to sequence my DNase-seq samples to perform digital footprinting?'. In order to assess the affect of sequencing depth on footprint detection, using the ENCODE SkMC DNase-seq dataset (owing to its large 550 million read depth), we randomly subsampled reads to simulate the effect of differing read depths. We then ran Wellington on these data, utilising an FDR of 0.01 to select footprints at the varying read depths. In line with the subsampling efforts performed by ENCODE (Neph et al., 2012); footprint detection of human DNase-seq data by these methods are currently limited by the number of sequencing reads, with a positive correlation between sequencing depth and number of footprints detected (Figure S2). The current answer to this question, is therefore 'as much as possible', as we have not yet reached a point where contributing more sequencing depth does not increase the number of footprints detected.

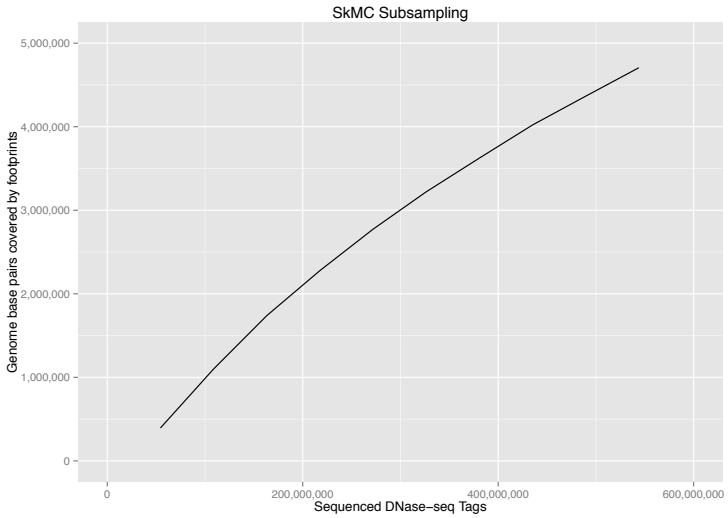


Figure S2: The number of footprints identified by Wellington in the SkMC DNase-seq dataset is positively correlated with sequencing depth. DNase-seq data were randomly subsampled in order to simulate different read depths of a single DNase-seq library, and the Wellington algorithm was used to identify footprints (FDR: 0.01).

1.10 Interpretation of the footprinting results

As mentioned before, note that p-values only give an indication of the strength of evidence against the null hypothesis, but do not provide a measure of footprint quality or strength. Further note that footprinting results depend on both the choice of significance threshold and possible values for l_{FP} and l_{SH}. We recommend trying Wellington with different parameters and observing how the choices influence the footprinting results as, in some situations, changing these can cause individual footprints to change their length or move by several base pairs. Therefore, when interpreting the footprinting results we need to be mindful of not to over-interpret minor differences in the p-values. In particular, when comparing footprints to known motifs, we recommend not requiring a 100% overlap between footprints and motifs, but some tolerance. As always, sensitivity analyses should be performed to see how any inferences made from the result depend on choices made.

1.11 Options to increase computational efficiency

There are several options to reduce computation time. The biggest gains can be achieved by restricting the search for footprints to regions of interest, such as DNase I hypersensitive sites (DHS). It has been established in the literature

that the vast majority of protein binding sites are in these regions (Cockerill, 2011). The pyDNase package can incorporate DHS coordinate information and restrict the footprint search to these regions without needing to load data from other parts of the genome. We have compared the footprinting results for ENCODE datasets for the whole genome and for DHS only and the differences in the results were small (data not shown). cl Computation time can also be reduced by reducing the number of possible values for l_{FP} and l_{SH}. For ENCODE datasets, we have found that restricting these to multiples of two does not significantly change the results.

The pyDNase package offer additional options which change how the algorithm processes data to accommodate certain computer setups (e.g. low amounts of RAM), but does not alter the algorithm nor impact the results. These are explained in the documentation provided with the package.

1.12 Alignability of the genome

We note that previous methods have included the consideration of the alignability of the target genome in their model (Hesselberth et al., 2009), whilst others do not (Pique-Regi et al., 2011; Boyle et al., 2011). In particular, a short unalignable region in between two alignable regions might be falsely identified as a footprint if the mappability of the genome is not taken into consideration. We initially allowed for mappability correction using a previous method (Hesselberth et al., 2009) but the results did not yield an increase in performance and increased compute time by a factor of 5-10x, extending analysis times to several days. The negligible difference in performance was because this hypothetical situation of short, unalignable regions in the genome is not particularly common, especially as read length increases beyond 36bp. Based on this, we decided in our final model that mappability would not be considered. We recommend that anyone wishing to perform mappability correction filters footprints that are in unmappable regions using their criterion of choice after footprint detection. This method has been utilized previously (Neph et al., 2012), who note that less than 1% of their footprints satisfy this criterion.

1.13 Possible extensions to the Wellington method

The Wellington method can be easily extended in several directions. The current binomial hypothesis test can be changed to a more complex null hypothesis that takes sequencing artefacts (spikes) in the data into account. Priors on

footprint lengths or even fully Bayesian approaches can easily be implemented by simply changing how the p-values are calculated. Additional preprocessing methods can also easily be added, e.g. to allow taking DNase I cutting preferences into account.

It remains to be seen how digital genomic footprinting can be used to compare multiple datasets covering differing states (e.g. healthy vs. diseased). The presence of a DHS in two cell types does not mean that the same event is occurring, as different transcription factors could be binding in a context sensitive manner. More detailed analyses into the differences in DNase cuts in DHSs between datasets would give an insight into the occupancy of promoters and enhancers by different transcription factors in differing states.

2 pyDNase

We noticed the need for a simple tool to handle DNase-seq data from a standard format, and to be able to access the data in a random order, as 90% of the genome is not hypersensitive and therefore processing these regions increases computational time and resources tenfold. pyDNase solves these problems by providing a simple Python interface to access cut information stored in the Binary Standard Alignment/Map (BAM) file format produced by popular mapping tools such as Bowtie and BWA and adopted by the ENCODE consortium as the preferred alignment file format.

pyDNase capitalises on the recent introduction of the SAM format to randomly access cut information in any region in the genome without the need to load the entire dataset at once. By using peak detection software such as FindPeaks, HOMER, Maq, HotSpots, etc to locate DNase Hypersensitive sites, footprinting algorithms can be performed solely on regions of interest, speeding up computational time dramatically.

Briefly, pyDNase uses a key-value array read cache which can be enabled or disabled at run time. If enabled, when DNase-seq cut data are requested from a genomic location for the first time, the surrounding 1000bp (configurable) will be automatically stored in memory for subsequent access. A key-value array is used as a sparse vector to store this data, as most sequences in the genome are not hypersensitive (and therefore have a data value of 0) so the use of spare data storage significantly reduces the memory footprint. Around 4GB of RAM is required to cache all the information in Human DNase I hypersensitive sites at once.

Full documentation and description of the features can be found at

`http://jpiper.github.com/pyDNase`

3 Validation of Footprints

3.1 ChIP-seq data

We used 'optimal' ChIP-seq peaks downloaded directly from the EBI ENCODE analysis FTP server. Names of the files along with summary statistics for each ChIP-seq experiment can be found in Table S1.

				ChIP-Seq Peaks			Genomic Motifs	
Cell Type	Track Name	ChIP factor	Total	With Motif	Without Motif	Inside Peaks	Outside Peaks
K562	EncodeHaibTfbsK562Atf3V0416101	ATF3	16,011	2,062	13,849	4,298	160,476
K562	EncodeSydhTfbsK562CmycStd	cMyc	5,023	2,098	2,925	4,311	509,474
K562	EncodeOpenChromChipK562Ctcf	CTCF	56,058	25,788	30,270	26,432	41,171
K562	EncodeUchicagoTfbsK562EjundControl	JunD	26,674	2,600	24,074	5,070	112,080
K562	EncodeHaibTfbsK562MaxV0416102	Max	46,171	16,419	29,752	34,226	1,131,669
K562	EncodeSydhTfbsK562Nfe2Std	NFE2	2,637	1,619	1,018	1,750	50,360
K562	EncodeSydhTfbsK562Nrf1Iggrab	NRF1	4,211	2,609	1,602	5,960	20,440
K562	EncodeHaibTfbsK562NrsfV0416102	NRSF	15,849	2,055	13,794	2,112	2,750
K562	EncodeHaibTfbsK562Pu1Pcr1x	PU.1	28,677	18,514	10,163	20,262	549,330
K562	EncodeHaibTfbsK562Sp1Pcr1x	Sp1	7,206	2,830	4,376	4,861	137,047
K562	EncodeHaibTfbsK562Usf1V0416101	USF1	18,521	12,431	6,090	23,808	524,900
A549	EncodeHaibTfbsA549Atf3V0422111Etoh02	ATF3	6,580	308	6,272	636	164,138
A549	EncodeSydhTfbsA549Bhlhe40Iggrab	bHLHE40	3,123	1,225	1,898	2,667	254,108
A549	EncodeSydhTfbsA549CebpbIggrab	CEBP	38,845	25,305	13,540	46,517	1,722,853
A549	EncodeUwTfbsA549CtcfStd	CTCF	45,732	23,536	22,196	24,289	43,314
A549	EncodeHaibTfbsA549Elf1V0422111Etoh02	ELF1	8,611	5,075	3,536	6,937	348,645
A549	EncodeHaibTfbsA549Ets1V0422111Etoh02	ETS1	5,525	2,564	2,961	3,466	1,145,432
A549	EncodeHaibTfbsA549GabpV0422111Etoh02	GABP	12,348	7,196	5,152	9,396	871,724
A549	EncodeSydhTfbsA549MaxIggrab	Max	9,881	3,982	5,899	8,965	1,156,930
A549	EncodeHaibTfbsA549NrsfV0422111Etoh02	NRSF	11,970	1,938	10,032	1,861	3,001
A549	EncodeHaibTfbsA549Usf1V0422111Etoh02	USF1	8,004	4,710	3,294	9,452	539,256
A549	EncodeHaibTfbsA549Yy1cV0422111Etoh02	YY1	10,259	2,148	8,111	2,079	52,874
A549	EncodeHaibTfbsA549Zbtb33V0422111Etoh02	ZBTB33	7,152	626	6,526	1,052	14,443

HepG2	EncodeHaibTfbsHepg2Atf3V0416101	ATF3	3,291	1,132	2,159	2,392	162,382
HepG2	EncodeOpenChromChipHepg2Cmyc	c-Myc	4,413	1,762	2,651	3,558	510,247
HepG2	EncodeHaibTfbsHepg2Ctcfsc5916V0416101	CTCF	55,778	26,856	28,922	27,655	39,948
HepG2	EncodeHaibTfbsHepg2Foxa2sc6554V0416101	FOXA1	40,989	29,356	11,633	76,105	6,363,320
HepG2	EncodeHaibTfbsHepg2Hnf4asc8987V0416101	HNF4a	20,805	10,913	9,892	12,889	519,231
HepG2	EncodeHaibTfbsHepg2JundPcr1x	JunD	21,614	866	20,748	1,632	115,518
HepG2	EncodeSydhTfbsHepg2MaxIggrab	Max	11,854	4,707	7,147	10,726	1,155,169
HepG2	EncodeHaibTfbsHepg2Mybl2sc81192V0422111	MYB	17,898	8,016	9,882	10,306	2,389,517
HepG2	EncodeSydhTfbsHepg2Nrf1Iggrab	NRF1	1,902	1,635	267	4,132	22,268
HepG2	EncodeHaibTfbsHepg2NrsfV0416101	NRSF	12,828	1,686	11,142	1,743	3,119
HepG2	EncodeHaibTfbsHepg2RxraPcr1x	RXR	17,063	6,976	10,087	9,044	1,265,857
HepG2	EncodeHaibTfbsHepg2Sp1Pcr1x	Sp1	25,477	3,599	21,878	6,087	135,821
HepG2	EncodeSydhTfbsHepg2Srebp1InslnStd	Srebp1a	2,585	293	2,292	307	327,404
HepG2	EncodeSydhTfbsHepg2TbpIggrab	TBP	13,806	2,490	11,316	3,798	3,136,789
HepG2	EncodeSydhTfbsHepg2Tr4Ucd	TR4	2,953	660	2,293	836	88,253
HepG2	EncodeHaibTfbsHepg2Usf1Pcr1x	USF1	21,890	14,809	7,081	27,503	521,205

Table S1: Summary information of ChIP-seq data used in this study, along with statistics displaying the number of ChIP-seq peaks with and without the transcription factor's corresponding motif, and the number of the genomic motif occurrences that are within ChIP-seq peaks.

3.2 CENTIPEDE and ENCODE data preparation

All motif instances in the genome were located using HOMER, and DNase-seq cuts were exported into the custom data format required by CENTIPEDE using pyDNase. We verified that our implementation was working using the example data and results provided by the authors. CENTIPEDE was then run on each transcription factor. ENCODE (Neph et al., 2012) footprinting data were downloaded from the EBI ENCODE analysis FTP server, and sorted according to the score assigned to each footprint. Their footprints were extended 7bp in each direction as per their processing instructions. We found that when this step is omitted, the ENCODE (Neph et al., 2012) footprints are often too small to overlap motif instances and the performance drops drastically.

3.3 Definition of performance characteristics

In order to calculate the Receiver Operator Characteristic (ROC), we must define a gold standard set of bound and unbound motifs using each ChIP-seq experiment. Using ChIP-seq derived matrices provided as part of the HOMER suite (Heinz et al., 2010), we searched the entire genome for known binding motif instances. Motifs which were found inside ChIP-seq peaks were said to be bound by its corresponding factor family, and motifs falling outside the ChIP-peaks, were considered to be unbound by its corresponding factor.

We then calculated the footprint predictions over a range of p-value thresholds for Wellington, the full range (0 to 0.95) of Footprint occupancy scores for ENCODE (Neph et al., 2012), and full range (0 to 1) Log-odds probabilities for CENTIPEDE, limiting our analyses to the DHSs provided by ENCODE. Using the same set of motif locations outlined above, we then used the Wellington footprints to split all genomic instances of the motifs into 'Predicted to be Bound' or 'Not Predicted to be Bound', if either 70% of the motif was contained within the footprint, or vice versa. Thus, we end up with the following classifications for ROC analysis.

- True Positives (TPs): Motif instances falling within ChIP-seq peaks that are correctly predicted as being bound by Wellington.
- True Negatives (TNs): Motif instances falling outside of ChIP-seq peaks that are correctly predicted as being unbound by Wellington.
- False Positives (FPs): Motif instances falling outside of ChIP-seq peaks that are incorrectly predicted as being bound by Wellington.

- False Negatives (FNs): Motif instances falling within ChIP-seq peaks that are incorrectly predicted as being unbound by Wellington.

This can either be measured on a site-wise basis, or base pair basis. In practice, we found little difference between the statistics in the results presented, but used the per base pair prediction statistic.

3.4 Performance statistics

Performance statistics were calculated as follows.

- Positive Predictive Value (PPV): $TP/(TP + FP)$
- Sensitivity (Coverage): $TP/(TP + FN)$
- False Positive Rate (FPR): $FP/(FP + TN)$
- Performance Coefficient (Tompa et al., 2005) (PPC): $TP/(TP + FN + FP)$

3.5 Conservation and motif content

Conservation was calculated by summing the Vertebrate phyloP46way values for each co-ordinate in a set of footprints, and then dividing by the number of basepairs to yield the average conservation per bp. Motif content was calculated by searching for motifs (using HOMER's ChIP-seq derived matrices) and then counting the number of basepairs in the predicted footprints that are overlapped by a motif (multiple overlapping motifs at one base pair do not increase the score), and then dividing by the number of basepairs to yield the average motifs per bp. *de novo* motif searching was performed using HOMER (Heinz et al., 2010).

4 Supplemental Figures

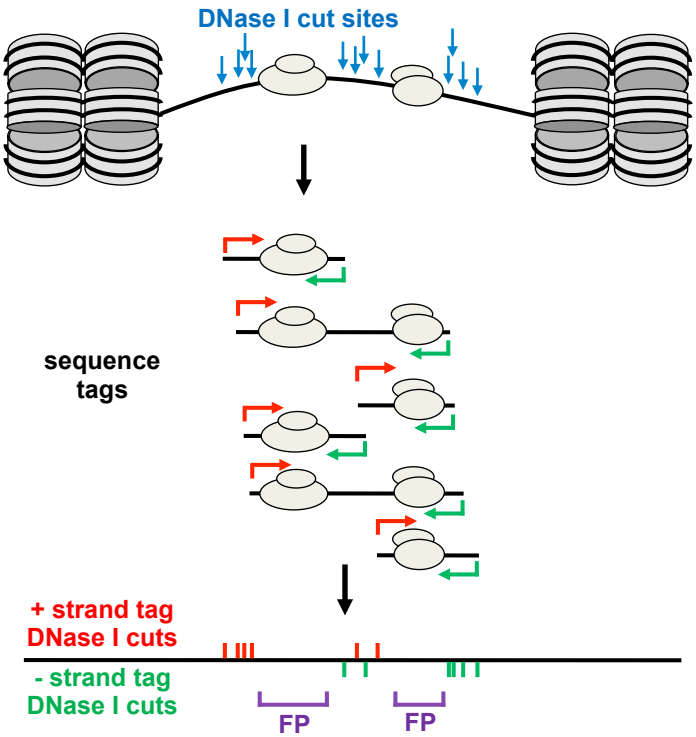
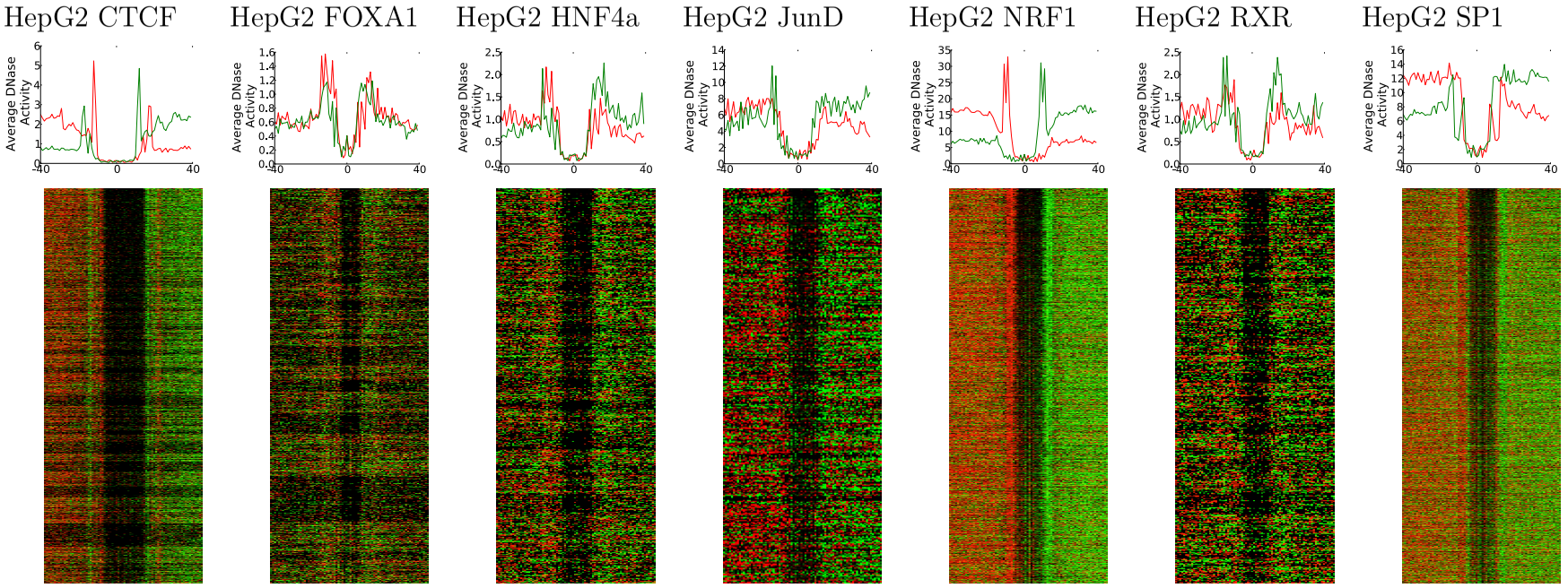


Figure S3: Chromatin structure based modelling of strand-specific DNase-seq data arising from DHSs. DHSs are usually 200-250 bp across, and the DNA sub-fragments of DHSs detected by DNase-seq are typically in the order of 50 to 150 bp in length and are surrounded by nucleosomal DNA. As depicted above, most of these fragments are expected to originate from within the DHS, meaning that they are likely span the regions of DNA protected by bound factors (indicated as ovals) that give rise to DNase I footprints (FPs). This means that it is the cut site that must be used to identify FPs, and not the entire sequence tag as is used in most peak detection algorithms. Furthermore, because sequence tags represent just one end of these fragments, upper strand +ve sequence data (red arrows) should represent sequences starting upstream of these FPs, while lower strand –ve sequence data (green arrows) should represent sequences starting downstream of FPs. The Wellington program has taken advantage of the fact that this strand information can be used to greatly increase the power of FP detection algorithms by making use of both the precise position of the cut site, and the predicted orientation of these cuts relative to a bound factor. As represented below the model, when the sum of the DNase I cuts in a DHS is depicted it is immediately apparent that FPs will generate a concentration of upstream +ve strand tags and a concentration of –ve strand downstream tags.


HepG2 CTCF
HepG2 FOXA1
HepG2 HNF4a
HepG2 JunD
HepG2 NRF1
HepG2 RXR
HepG2 SP1
Average DNase Activity

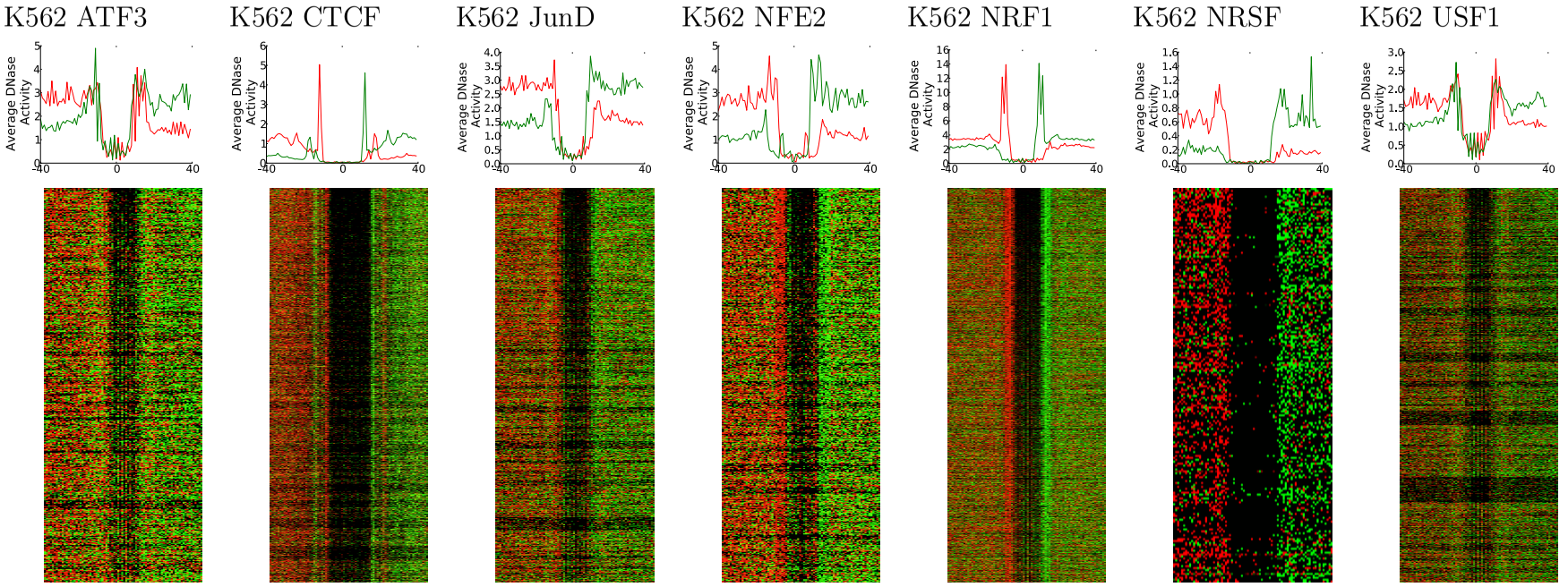


Figure S4: DNase-seq cleavage imbalance in double-hit DNase-seq data is observed at multiple transcription factor binding sites and multiple cell types. Upper panels: DNase-seq cleavage patterns surrounding ChIP-seq verified binding sites in HepG2 and K562 cells with a Footprint Occupancy Score (FOS) (Neph et al., 2012) of <0.95 illustrate the abundance of sequencing fragments aligning to the positive reference strand (red) upstream of protein-DNA binding sites, and to the negative reference strand (green) downstream of the protein-DNA binding site. Note cell-type independent differences and similarities between cleavage patterns for each transcription factor family. Lower panels: Heat map depicting cutting frequencies around the binding site. Binding sites are sorted from top to bottom in order of decreasing FOS .

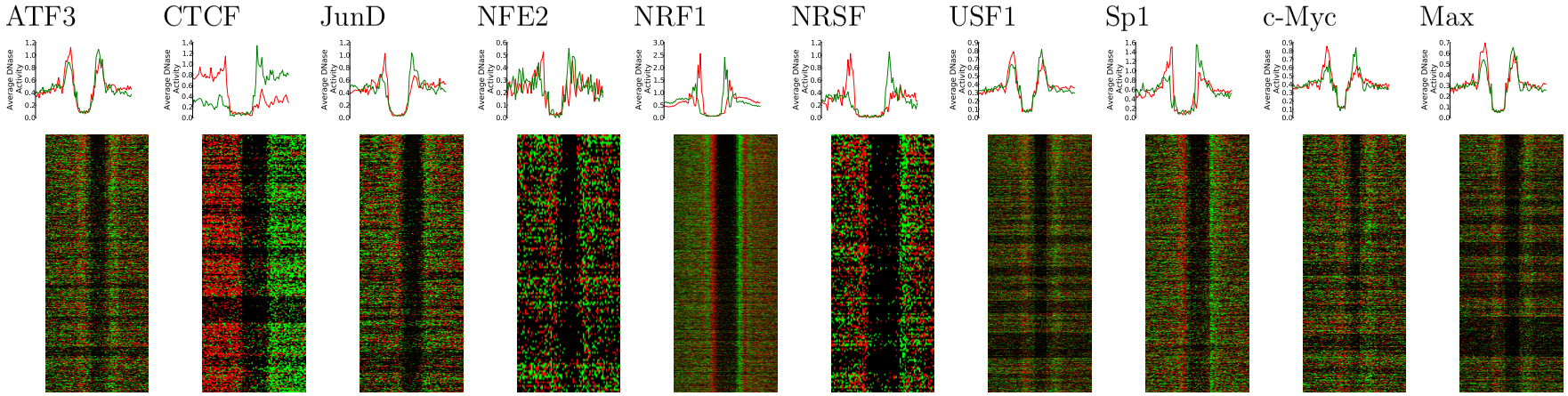


Figure S5: DNase-seq cleavage imbalance is less pronounced in DNase-seq data generated using the original single-hit DNase-seq library preparation protocol. Upper panels: DNase-seq cleavage patterns surrounding ChIP-seq verified binding sites in K562 cells with a Footprint Occupancy Score (FOS) of <0.95 illustrate the abundance of sequencing fragments aligning to the positive reference strand (red), and to the negative reference strand (green). Some transcription factors (CTCF, NRF1, NRSF) demonstrate cleavage imbalance patterns consistent with the double-hit protocol (Figure S4), whereas others (ATF3, JunD, NFE2) exhibit diminished strand imbalance. Lower panels: Heat map depicting cutting frequencies around the binding site. Binding sites are sorted from top to bottom in order of decreasing FOS.

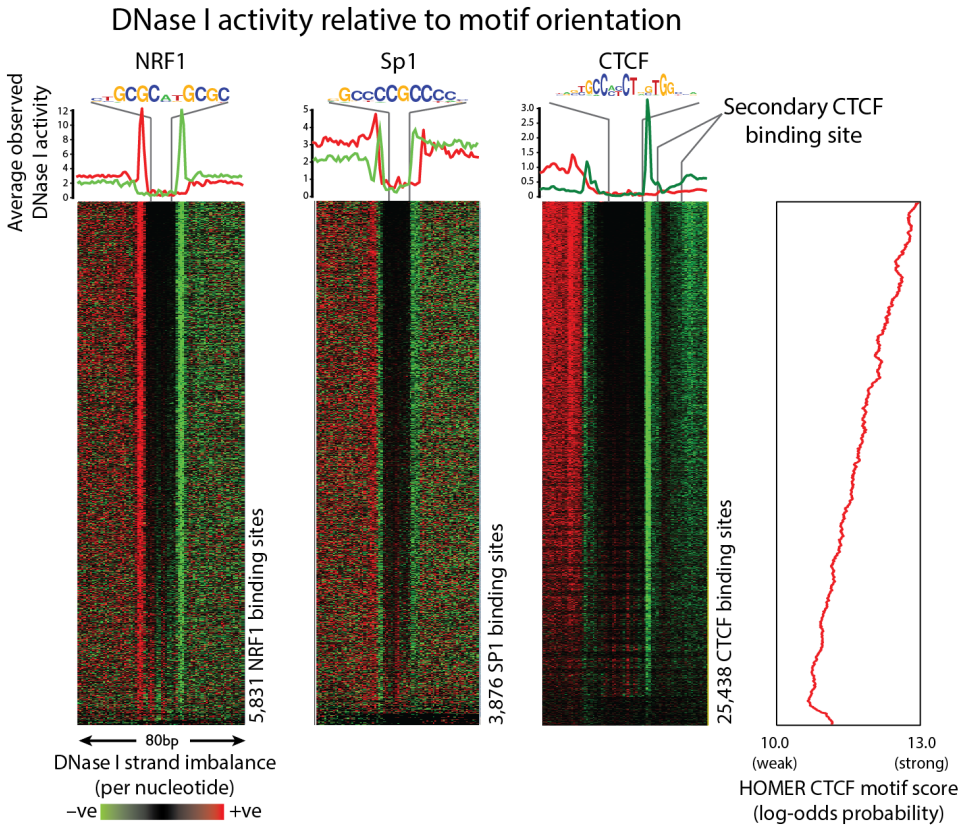


Figure S6: Heat maps show transcription factor specific DNase-seq cleavage patterns surrounding verified NRF1, Sp1, and CTCF binding sites. Here the data are oriented relative to motif strand which is indicated in the upper panels. Red indicates an excess of positive strand (with respect to motif strand) cuts over negative strand (with respect to motif strand) cuts per nucleotide position, and green indicates an excess of negative strand (with respect to motif strand) cuts. Binding sites are sorted from top to bottom in order of decreasing Footprint Occupancy Score (Neph et al., 2012). Note that alignment of the CTCF motifs in the same orientation reveals an additional region corresponding to a secondary motif for CTCF binding (resembling CTGCAG; (Bowers et al., 2009; Boyle et al., 2011)) that is also protected. For CTCF, the HOMER motif scores are presented as a moving average on the right, highlighting the fact that decreasing footprint scores equate with decreasing motif scores. Additional sub-patterns of DNase I cleavage can also be seen within motifs with lower occupancy scores, perhaps reflecting inefficient binding of some specific individual DNA-binding domains when bound to sub-optimal motifs.

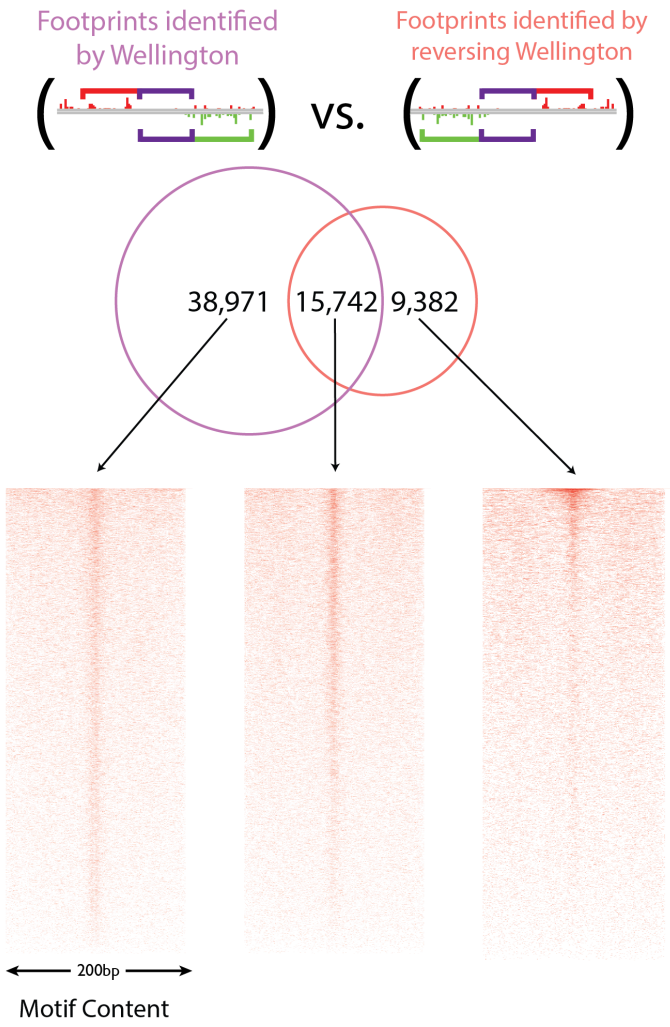


Figure S7: Motif content and motif location of Wellington and reverse Wellington footprints. Heatmaps of motif locations surrounding footprints identified by Wellington and reverse Wellington demonstrates the depletion of motifs at the centre of reverse Wellington footprints.

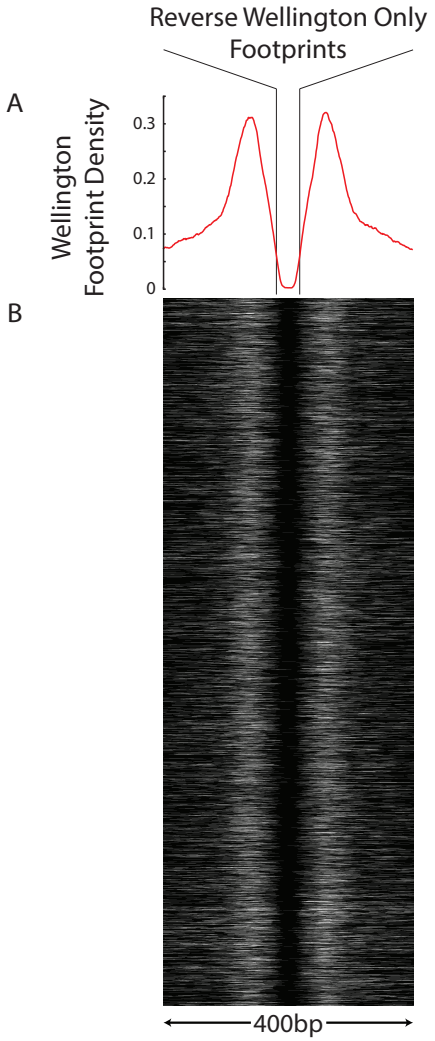
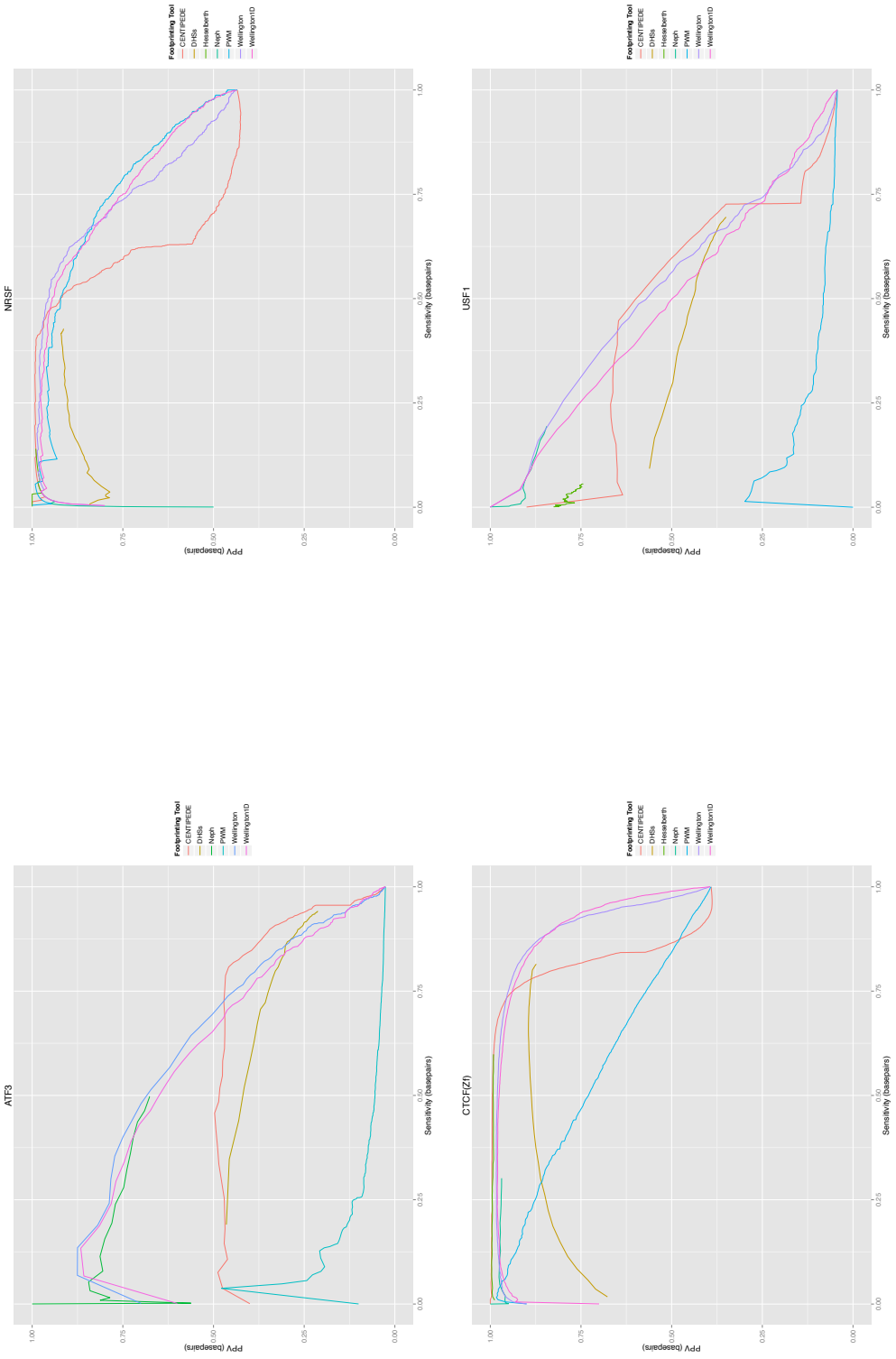
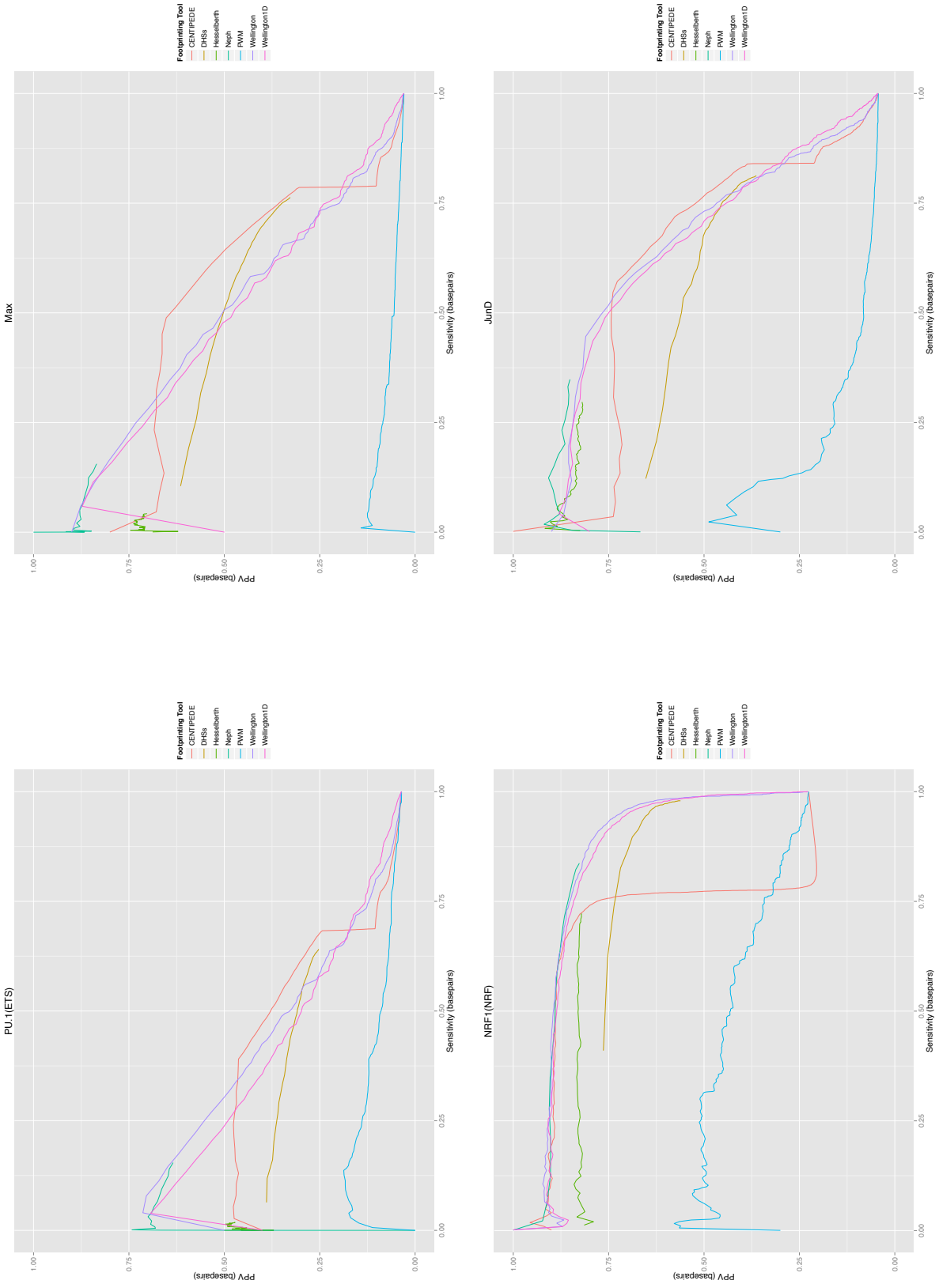


Figure S8: The majority of the 9,382 false positive footprints identified only by Reverse Wellington are located adjacent to or inbetween footprints identified by Wellington. (A) The distribution of Wellington footprints surrounding the 9,382 Reverse Wellington footprints, shown as the percentage of nucleotides at this position surrounding a Reverse Wellington Footprint which are found in a Wellington Footprint. (B) Heat map of footprints identified by Wellington centred on those only identified by Reverse wellington.


NFSF
USF1
ATF3
CTCF(Zf)
Footprinting Tool
CENTIPEDE
DHSs
Hexamerbeth
Neph
PWM
Wellington
Wellington1D
Sensitivity (basepairs)
PPV (basepairs)
1.00
0.75
0.50
0.25
0.00


Footprinting Tool
CENTIPEDE
DHSs
Hesselberth
Neph
PWM
Wellington
Wellington1D
Max
JunD
PU.1(ETS)
NRF1(NRF)
PPV (basepairs)
Sensitivity (basepairs)
1.00
0.75
0.50
0.25
0.00

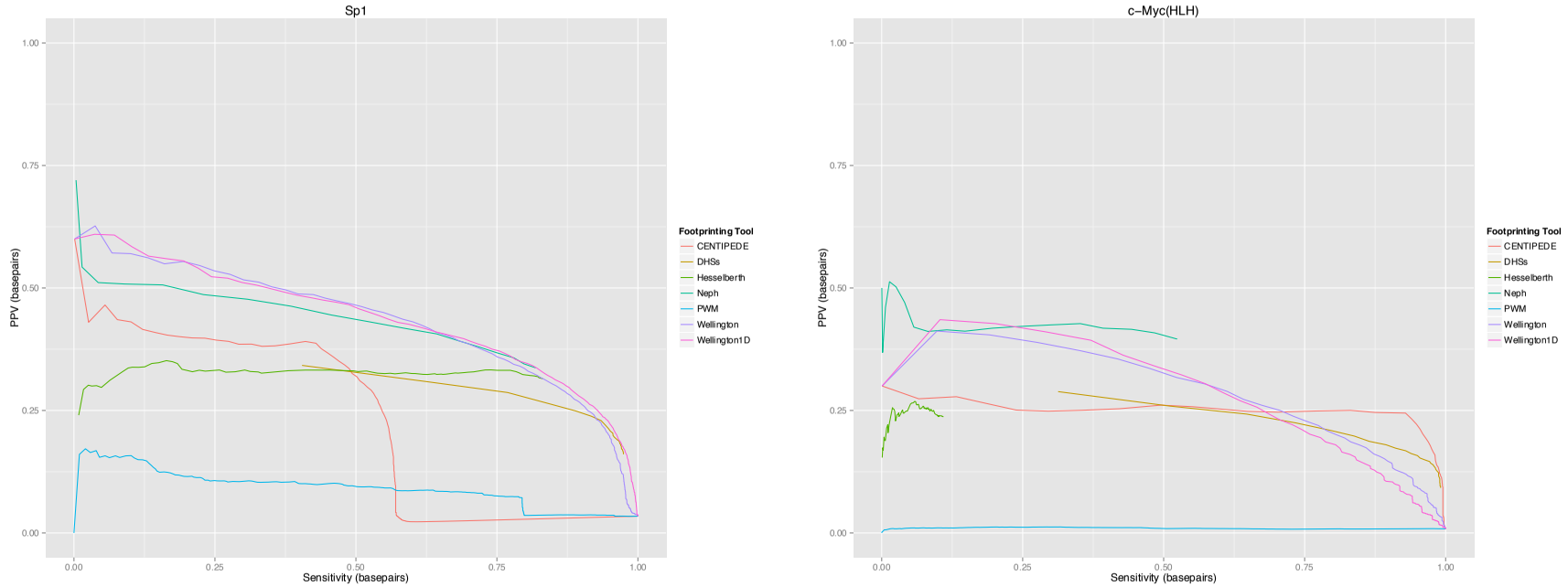


Figure S9: Positive Predictive Value of footprint predictions on ENCODE double-hit K562 DNase-seq data as a function of ChIP-seq sensitivity for 10 genomic transcription factor binding sites for Wellington, Wellington 1D, Neph et al., Hesselberth et al., Position Weight Matrices, DNase Hypersensitive Sites, and CENTIPEDE.

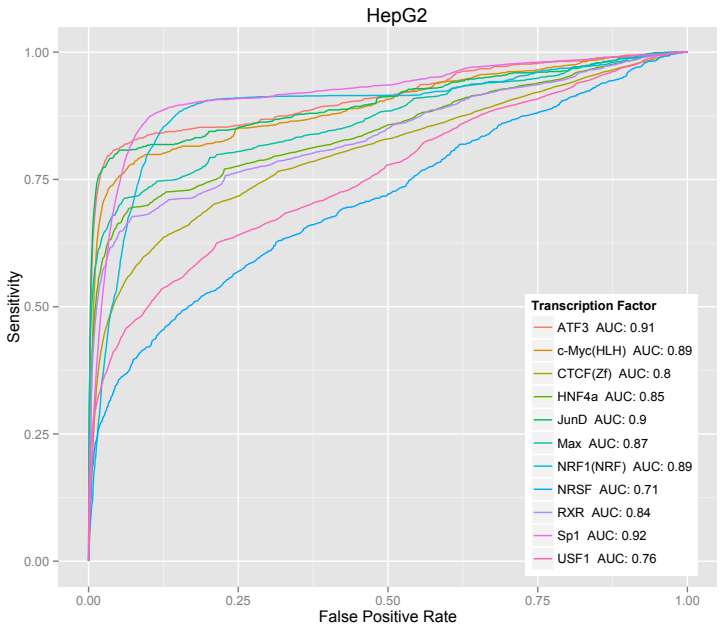


Figure S10: ROC analysis for 11 genomic transcription factor binding site predictions by Wellington using data from HepG2 cells

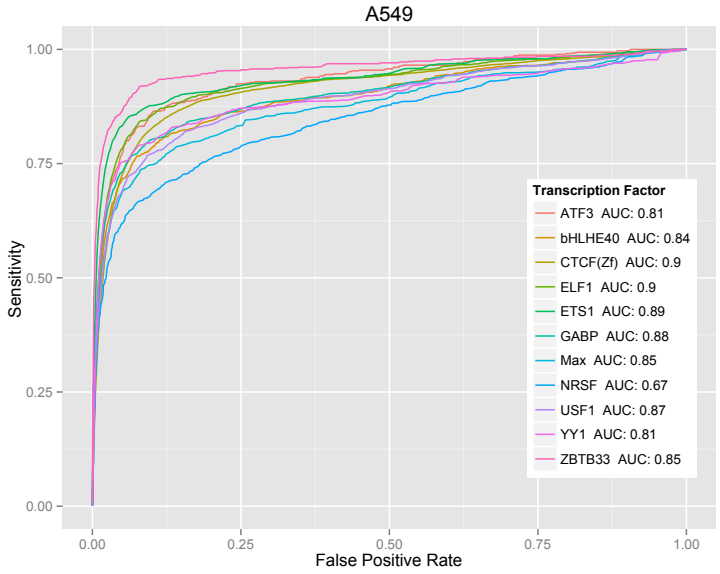


Figure S11: ROC analysis for 11 genomic transcription factor binding site predictions by Wellington using data from A549 cells

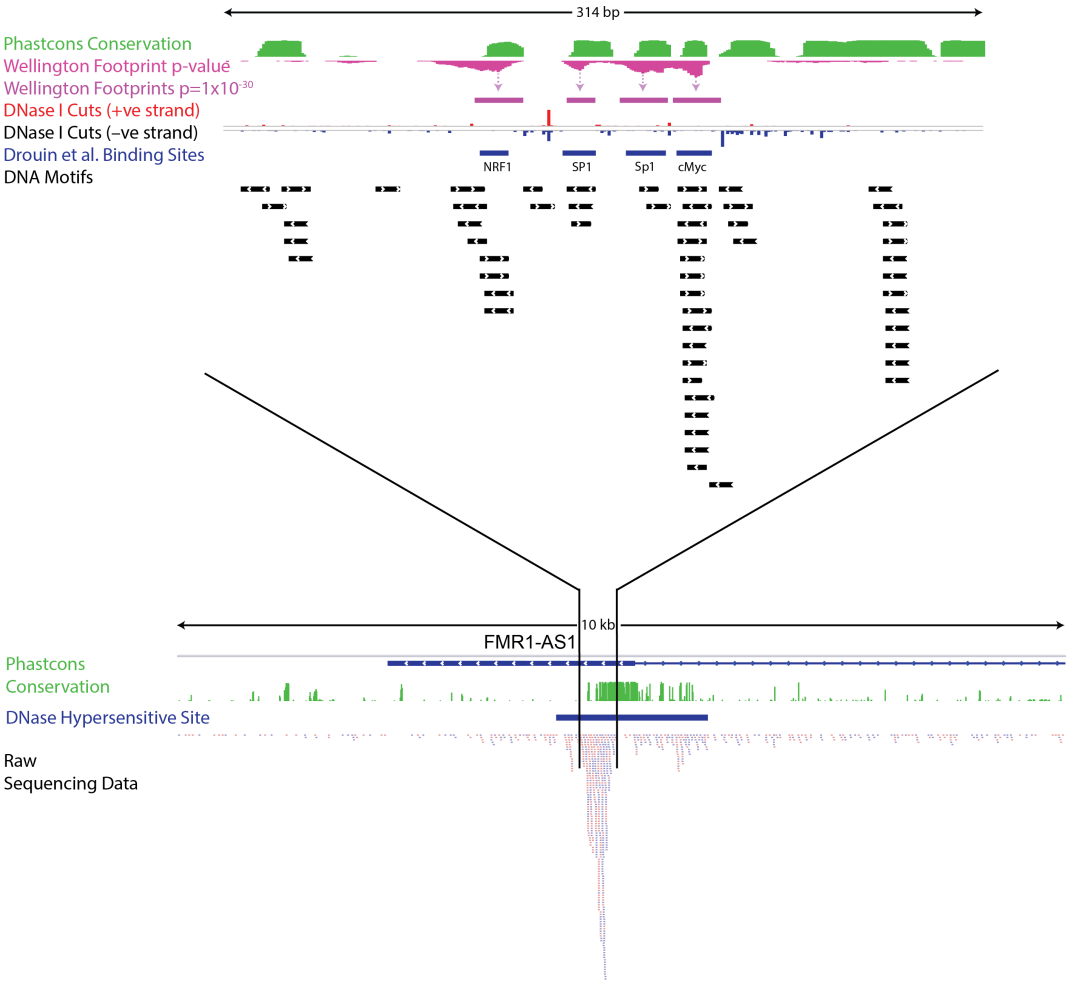


Figure S12: Footprints at the FMR1 promoter overlap with regions of high sequence conservation and improve over basic DNase-seq peak calling. Whilst phastcons conservation and DNA motifs overlap with known binding sites in this region, without footprinting, motif content alone is unable to predict bound locations. After applying the Wellington algorithm to the 1kb DNase hypersensitive site covering the FMR1 promoter, we produce footprints that align with the known protein-DNA interactions in this region without any off-target hits.

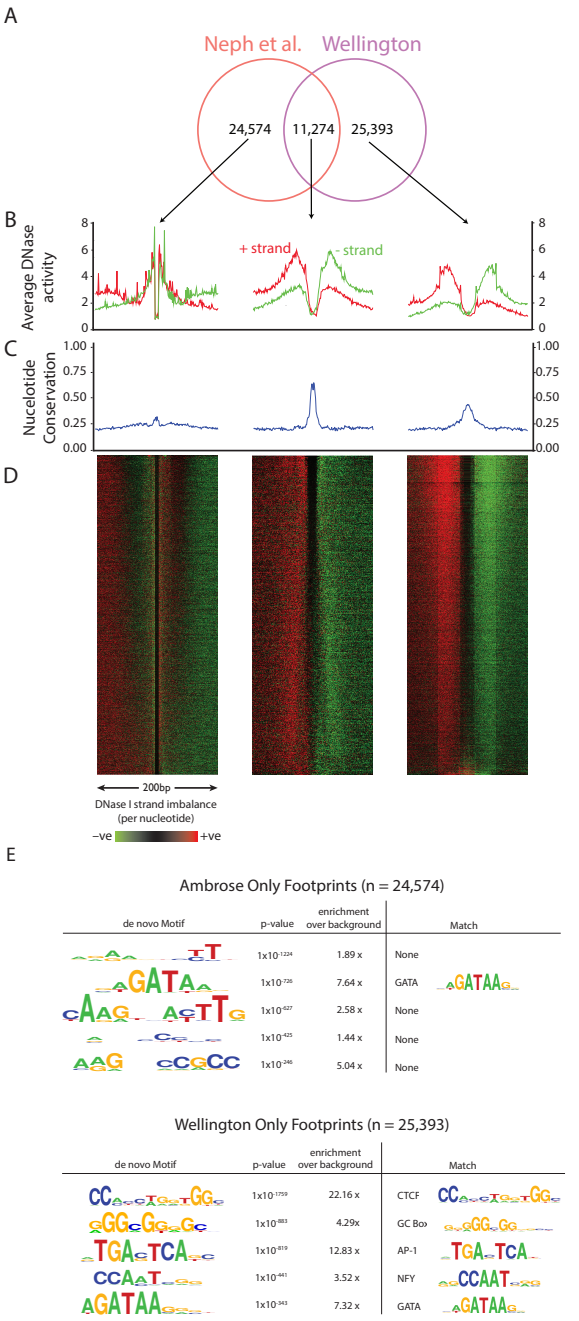


Figure S13: Footprints only identified by ENCODE (Neph et al., 2012) do not exhibit typical asymmetry. By comparing the 40,000 top scoring footprints for K562 cells from the ENCODE and the Wellington set, we observe that ENCODE exclusive footprints do not exhibit typical strand asymmetry identified in Figure 2 and have low average PhyloP conservation scores. *De novo* motif finding results show that Wellington footprints show more specific sequence logos, more enrichment over background, and more matches to known matrices.

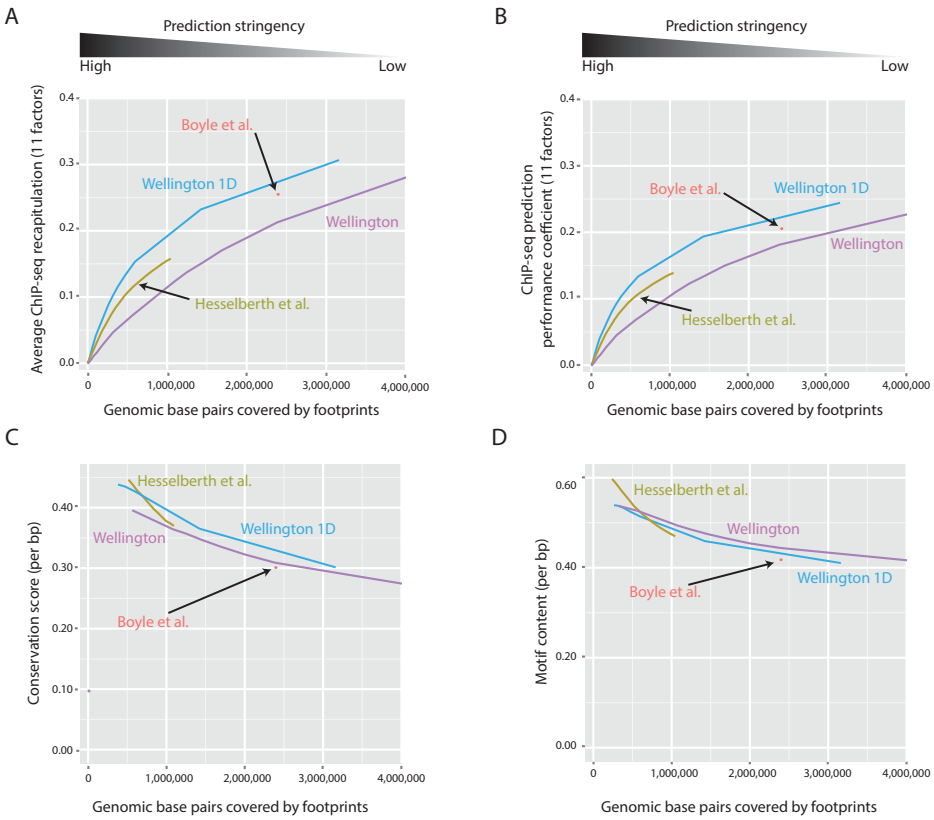
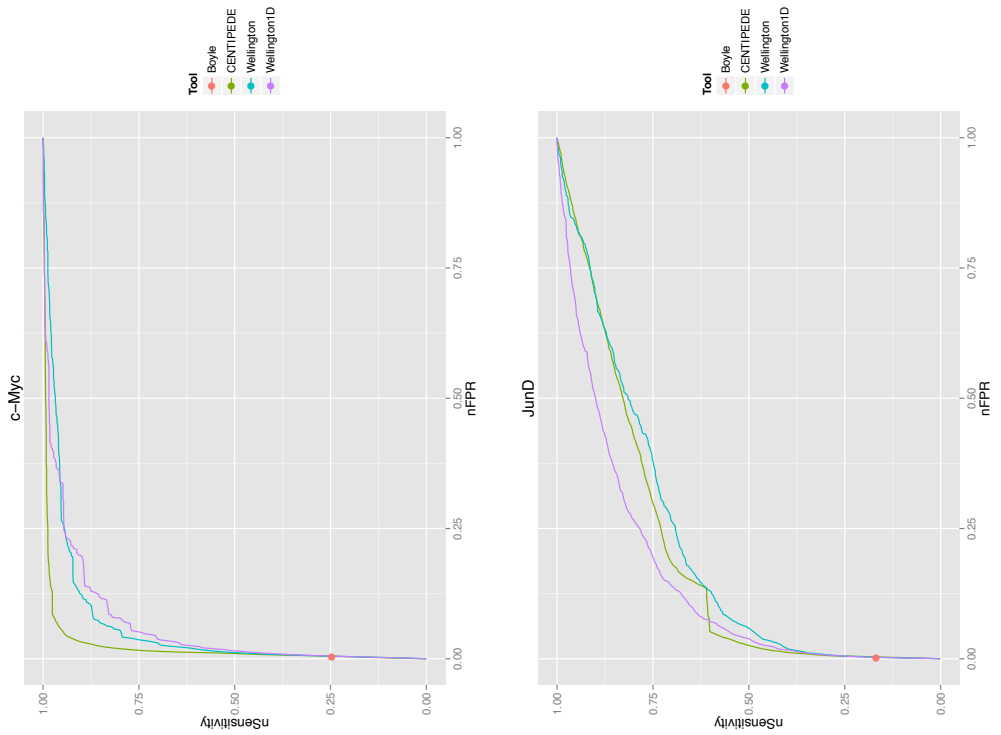
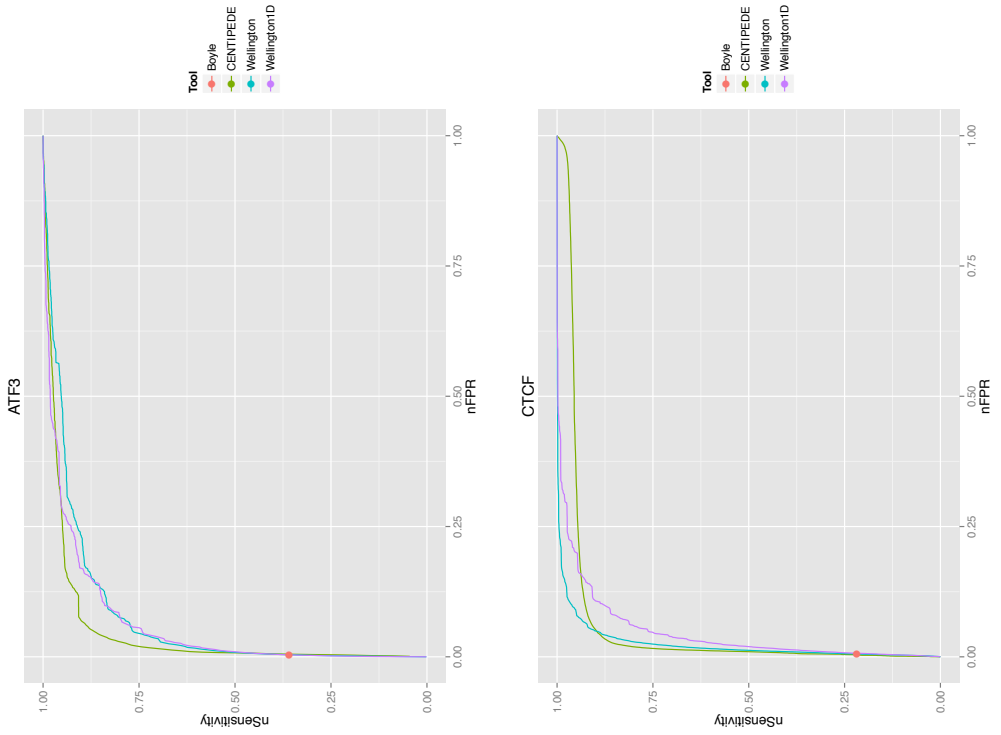
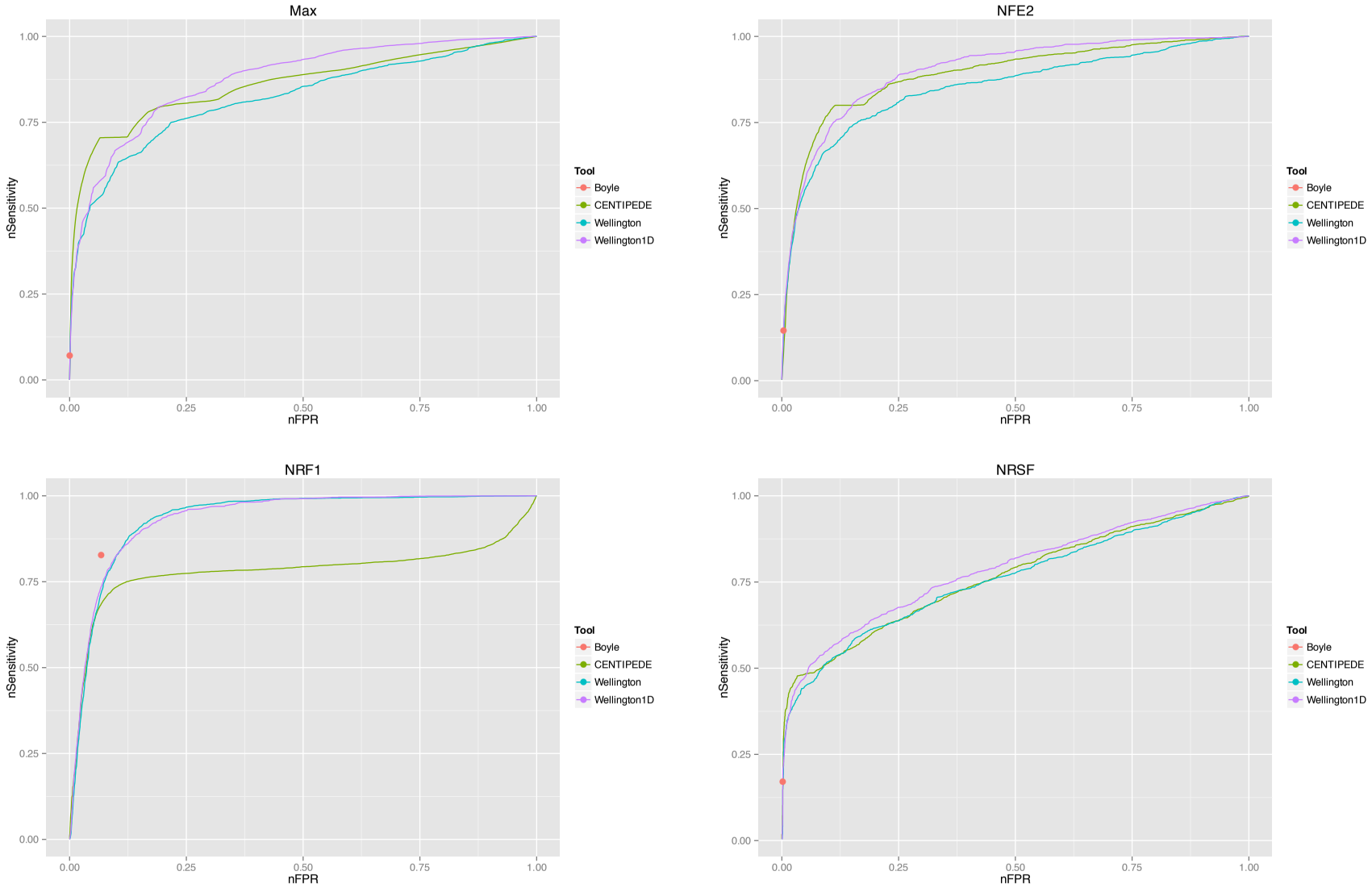


Figure S14: Wellington and Wellington 1D can also be used on DNase-seq data generated using the single-hit protocol. (A) Wellington is able to recapitulate a larger amount of ChIP-seq data than the predictions by Boyle et al. and Hesselberth et al. The horizontal axis shows the total number of base pairs in the genome that are covered by footprints at a given footprinting stringency, the vertical axis shows the average performance of these footprints in recapitulating binding sites found from ChIP-seq data for 11 transcription factors in K562 cells. (B) The nucleotide performance coefficients for these predictions Tompa:2005gx take numbers of false positives and false negatives into account and show a consistent finding compared to (A). (C, D) Wellington and Wellington 1D footprints have comparable conservation scores and motif content over a range of sensitivities, and with the available implementation, are able to detect more footprints than Hesselberth et al.


c−Myc
JunD
Tool
Boyle
CENTIPEDE
Wellington
Wellington1D
nSensitivity
nFPR
0.00
0.25
0.50
0.75
1.00


ATF3
CTCF
Tool
Boyle
CENTIPEDE
Wellington
Wellington1D
nSensitivity
nFPR
0.00
0.25
0.50
0.75
1.00


Max
NFE2
NRF1
NRSF
nSensitivity
nFPR
0.00
0.25
0.50
0.75
1.00
Tool
Boyle
CENTIPEDE
Wellington
Wellington1D

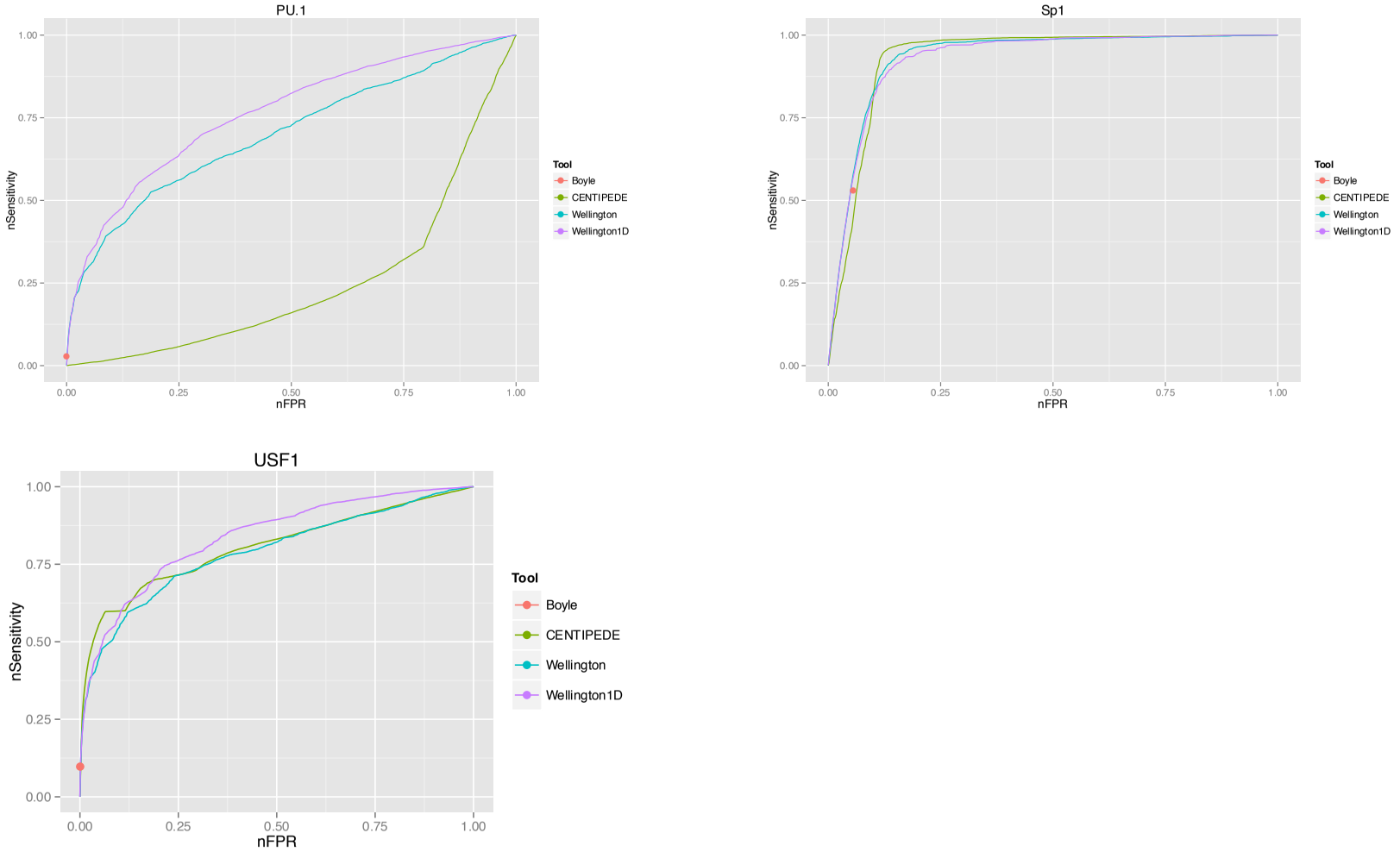
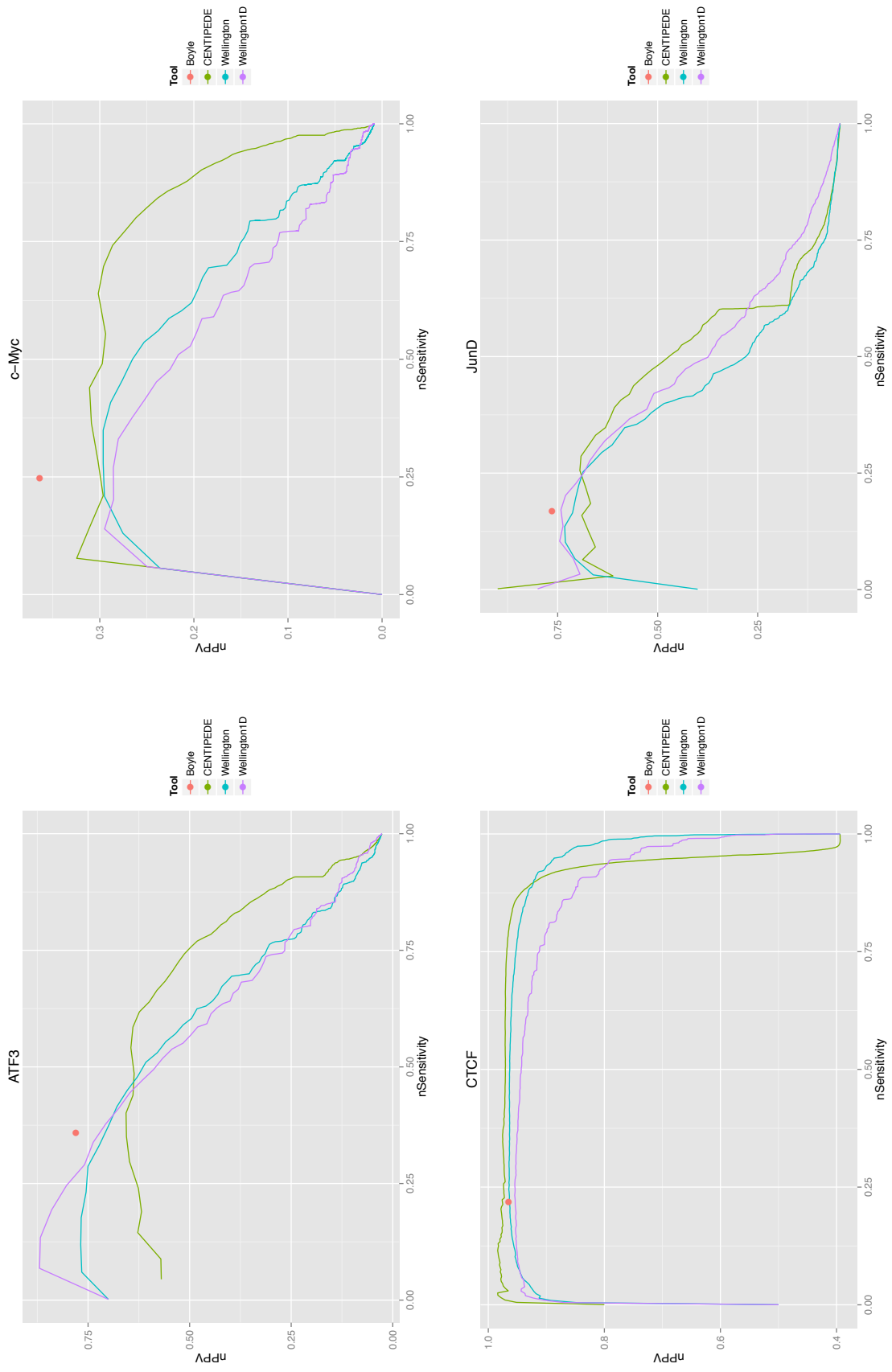
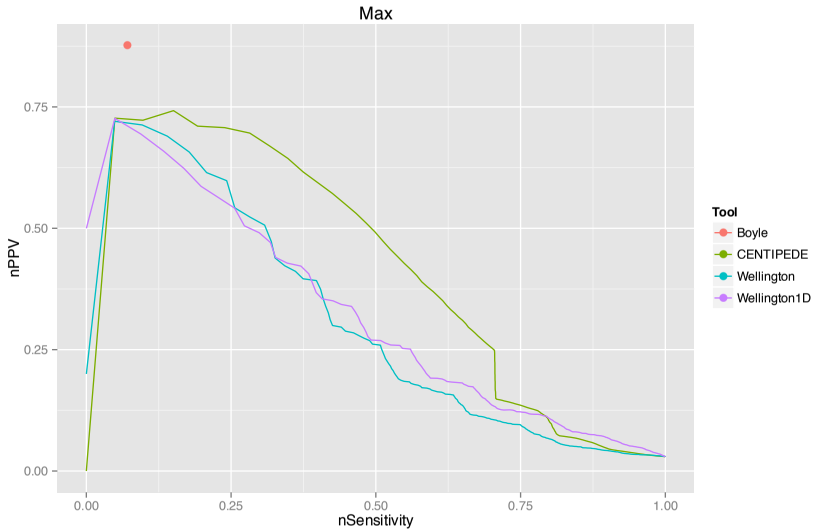
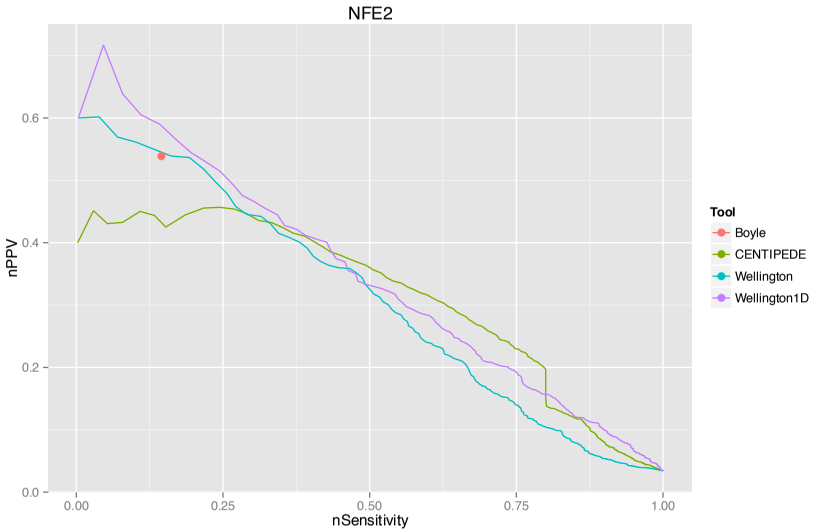
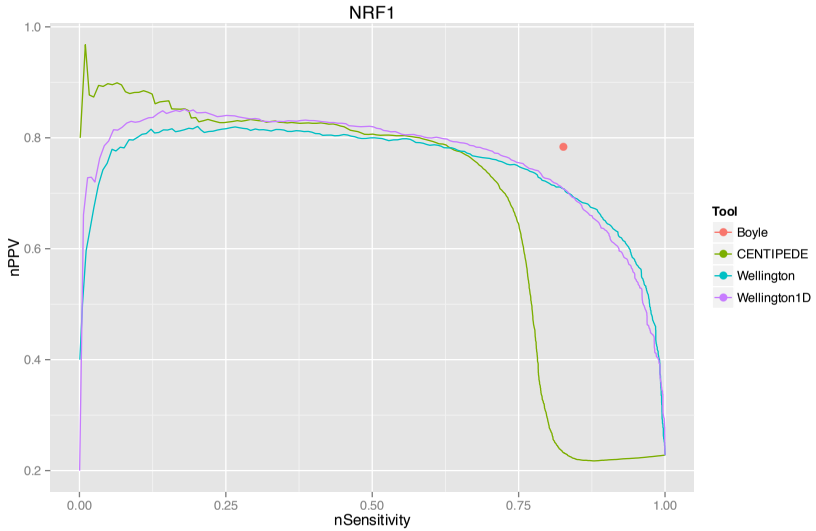


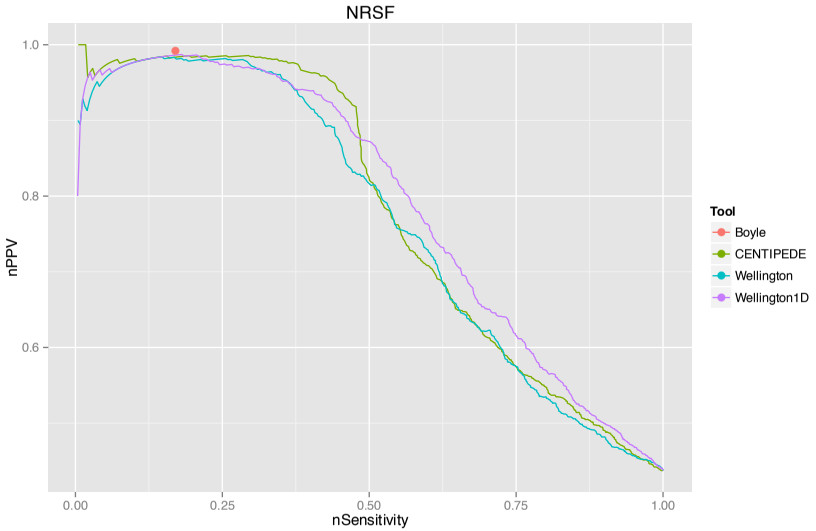
Figure S15: ROC analysis for Wellington, Wellington 1D, Boyle et al., and CENTIPEDE for transcription factor binding site predictions using K562 DNase-seq data data generated by the original single-hit library preparation.


c−Myc
JunD
ATF3
CTCF
nPPV
nSensitivity
Tool
Boyle
CENTIPEDE
Wellington
Wellington1D


Max
nPPV
nSensitivity
0.00
0.25
0.50
0.75
1.00
Tool
Boyle
CENTIPEDE
Wellington
Wellington1D


NFE2
nPPV
nSensitivity
0.0
0.2
0.4
0.6
0.00
0.25
0.50
0.75
1.00
Tool
Boyle
CENTIPEDE
Wellington
Wellington1D


NRF1
nPPV
nSensitivity
0.2
0.4
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00
Tool
Boyle
CENTIPEDE
Wellington
Wellington1D


NRSF
nPPV
nSensitivity
0.6
0.8
1.0
0.00
0.25
0.50
0.75
1.00
Tool
Boyle
CENTIPEDE
Wellington
Wellington1D

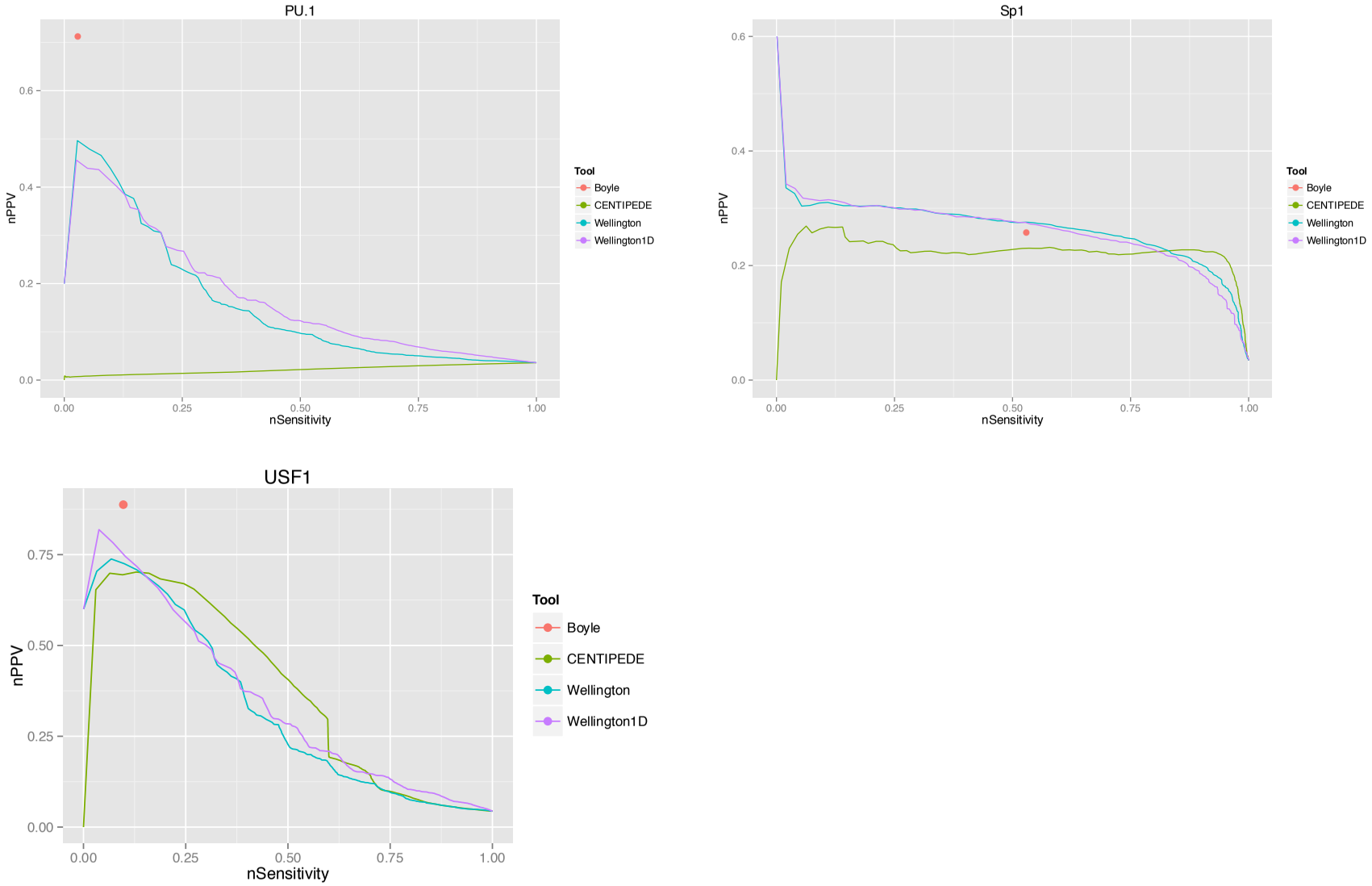


Figure S16: Positive Predictive Value of footprint predictions on K562 DNase-seq data data generated by the original single-hit library preparation as a function of ChIP-seq sensitivity for 11 genomic transcription factor binding for Wellington, Wellington 1D, Boyle et al. and CENTIPEDE.

References

S. R. Bowers, F. Mirabella, F. J. Calero-Nieto, S. Valeaux, S. Hadjur, E. W. Baxter, M. Merkenschlager, and P. N. Cockerill. A conserved insulator that recruits CTCF and cohesin exists between the closely related but divergently regulated interleukin-3 and granulocyte-macrophage colony-stimulating factor genes. *Molecular and cellular biology*, 29(7):1682–1693, Apr. 2009.

A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, Jan. 2008.

A. P. Boyle, L. Song, B.-K. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford, and T. S. Furey. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, 21(3):456–464, Mar. 2011.

P. N. Cockerill. Structure and function of active chromatin and DNase I hypersensitive sites. *The FEBS journal*, 278(13):2182–2210, July 2011.

S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*, 38(4):576–589, May 2010.

J. R. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields, and J. A. Stamatoyannopoulos. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, 6(4):283–289, Apr. 2009.

H. Koohy, T. A. Down, and T. J. Hubbard. Chromatin Accessibility Data Sets Show Bias Due to Sequence Specificity of the DNase I Enzyme. *PloS one*, 8 (7):e69853, 2013.

S. Neph, J. Vierstra, A. B. Stergachis, A. P. Reynolds, E. Haugen, B. Vernot, R. E. Thurman, S. John, R. Sandstrom, A. K. Johnson, M. T. Maurano, R. Humbert, E. Rynes, H. Wang, S. Vong, K. Lee, D. Bates, M. Diegel, V. Roach, D. Dunn, J. Neri, A. Schafer, R. S. Hansen, T. Kutyavin, E. Giste, M. Weaver, T. Canfield, P. Sabo, M. Zhang, G. Balasundaram, R. Byron,

M. J. MacCoss, J. M. Akey, M. A. Bender, M. Groudine, R. Kaul, and J. A. Stamatoyannopoulos. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, Sept. 2012.

R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3):447–455, Mar. 2011.

P. J. Sabo, M. S. Kuehn, R. Thurman, B. E. Johnson, E. M. Johnson, H. Cao, M. Yu, E. Rosenzweig, J. Goldy, A. Haydock, M. Weaver, A. Shafer, K. Lee, F. Neri, R. Humbert, M. A. Singer, T. A. Richmond, M. O. Dorschner, M. McArthur, M. Hawrylycz, R. D. Green, P. A. Navas, W. S. Noble, and J. A. Stamatoyannopoulos. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature Methods*, 3(7):511–518, July 2006.

M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Régnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, 23(1):137–144, Jan. 2005.

pyDNase Documentation

Jason Piper

September 01, 2013

Contents

1 Installation **2**
- 1.1 Supported systems . . . 2
- 1.2 Pre-installation requirements . . . 2
- 1.3 Installing `pyDNase` . . . 3

2 Getting DNase-seq cut data from BAM files **3**

3 Handling Genomic Intervals **4**
- 3.1 `GenomicInterval` . . . 4
- 3.2 `GenomicIntervalSet` . . . 5

4 Footprinting DNase-seq data **6**

5 Scripts **7**
- 5.1 example_footprint_scores.py . . . 7
- 5.2 dnase_average_profile.py . . . 8
- 5.3 dnase_wig_tracks.py . . . 9
- 5.4 dnase_to_javatreeview.py . . . 11
- 5.5 wellington_footprints.py . . . 12

6 Frequently Asked Questions **12**
- 6.1 How can I identify hypersensitive sites in DNase-seq data? . . . 13

Introduction Many people currently analyzing DNase-seq data are using tools designed for ChIP-seq work, but may be inappropriate for DNase-seq data where one is less interested in the overlaps of sequenced fragments, but the site at which the cut occurs (the 5' most end of the aligned sequence fragment).

We have developed *pyDNase* to interface with a sorted and indexed BAM file from a DNase-seq experiment, allowing efficient and easy random access of DNase-seq cut data from any genomic location, e.g.

```
>>> import pyDNase
>>> reads = pyDNase.BAMHandler(pyDNase.example_reads())
>>> reads["chr6,170863500,170863532,+"]
{'+': array([0,0,0,1,0,0,1,1,2,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,1,1,0,0,0,1]),
 '-': array([0,10,1,0,1,0,4,9,0,1,0,2,1,0,0,0,0,0,3,0,6,3,0,0,0,1,1,1,3,0,3,6])}
```

Querying the `BAMHandler` object returns a dictionary containing numpy arrays with DNase cut counts on the positive reference strand (+), and cuts on the negative reference strand (-). *pyDNase* efficiently caches the cut data queried, so that multiple requests from the same genomic locations do not require repeated lookups from the BAM file (this can be disabled).

pyDNase comes with several analysis scripts covering several common use cases of DNase-seq analysis, and also an implementation of the Wellington and Wellington 1D footprinting algorithms.

to install *pyDNase*, ensure NumPy is installed, and run:

```
$ pip install pyDNase
```

for full documentation go to: http://pythonhosted.org/pyDNase/

Support If you're having any troubles, please send an email to *j.piper@warwick.ac.uk* and I'll do my best to help you out. If you notice any bugs, then please raise an issue over at the github repo.

Contributions I highly encourage contributions! This is my first software development project - send any pull requests this way. I'm particularly interested in cool analysis scripts that anyone has written.

Reference

Note: If you use *pyDNase* or the Wellington algorithm in your work, please cite the following paper.

Piper et al. 2013. *Wellington: A novel method for the accurate identification of digital genomic footprints from DNase-seq data*, Nucleic Acids Research 2013; doi: 10.1093/nar/gkt850

Contents

1 Installation

1.1 Supported systems

Hardware 1GB of RAM and a 64-bit operating system. RAM usage heavily depends on what you're doing, but 1GB is the bare minimum if you disable caching.

Software Tested on OS X 10.8 and on Ubuntu 11.10, but should run fine on any other *NIX flavour as long as the prerequisites are fulfilled.

Note: Windows is not supported.

1.2 Pre-installation requirements

In order to install `pyDNase`, the following software is required. Most people will already have most of these on their system. I have attempted to list them in the order that you need to install them in.

1. **A compiler suite** You can check by opening up the terminal and typing

   ```
   $ clang --version
   ```

   or

   ```
   $ gcc --version
   ```

   As long as you get a response from one of these, you're good to go. Failing that...

   - **On OS X < 10.7.3:** Install "Xcode" from https://developer.apple.com/downloads/
   - **On OS X >= 10.7.3:** Install "Command Line Tools for Xcode" from https://developer.apple.com/downloads/

   (you can also install Xcode, but this is overkill) * **On Ubuntu:** Install with `sudo apt-get install build-essentials` * If you're using some other *NIX distro, I assume you know what you're doing.

2. **Python >= 2.6 (including Python 3!)**

   - This will come installed with OS X or any respectable *NIX distro.

3. **pip:** Used for automated installation of Python packages. If you don't already have pip installed, you can use the following command to install it

   ```
   $ curl https://raw.github.com/pypa/pip/master/contrib/get-pip.py | python
   ```

4. **Cython**

   - Provided you installed pip, you should be able to simply run

     ```
     $ pip install Cython
     ```

5. **samtools**

   - **On OS X** the simplest way to install samtools is using the homebrew command `brew tap homebrew/science`

   followed by `brew install homebrew/science/samtools`.

   - **On Ubuntu** you can use `sudo apt-get install samtools`

6. **NumPy**

   - Provided you installed pip, you should be able to simply run

     ```
     $ pip install numpy
     ```

1.3 Installing `pyDNase`

To install, simply

```
$ pip install pyDNase
```

This will attempt to download, compile, and install the python dependencies (`clint`, `numpy`, `scipy`, `pysam`, and `matplotlib`) automatically. However, due to a myriad of reasons it might not work. If this is the case, go and install these manually in said order, then try `pip install pyDNase` once more.

2 Getting DNase-seq cut data from BAM files

At the heart of the pyDNase package is the `BAMHandler` class, which provides an interfact to the cut data in a BAM file corresponding to a DNase-seq dataset. The interface is extremely simple:

```
>>> import pyDNase
>>> reads = pyDNase.BAMHandler("pyDNase/test/data/example.bam")
>>> reads["chr6,170863500,170863532,+"]
{'+': array([0,0,0,0,1,0,0,1,1,2,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,1,1,0,0,0]),
'-': array([0,10,1,0,1,0,4,9,0,1,0,2,1,0,0,0,0,0,3,0,6,3,0,0,0,1,1,1,3,0,3,6])}
```

As you can see, querying the `BAMHandler` object returns a dictionary containing numpy arrays with cut count on the positive reference strand (+), and cuts on the negative reference strand (-). If you wanted to look at the cuts with reference to something on the opposite strand, you can rotate the data 180 degrees by passing a "-" flag,

```
>>> reads["chr6,170863500,170863532,-"]
{'+': array([6,3,0,3,1,1,1,0,0,0,3,6,0,3,0,0,0,0,0,1,2,0,1,0,9,4,0,1,0,1,10,0]),
'-': array([0,0,0,1,1,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,2,1,1,0,0,1,0,0,0,0])}
```

By default, the BAMHandler caches lookups in 1000bp chunks. You can alter this behaviour at instanstiation. The BAMHandler also gives an interface to the Footprint Occupancy Score (FOS).

class `pyDNase.`**`BAMHandler`**(*filePath*, *caching=True*, *chunkSize=1000*)
The object that provides the interface to DNase-seq data help in a BAM file

FOS (*interval*, *bgsize=35*)
Calculates the Footprint Occupancy Score (FOS) for a Genomicinterval. See Neph et al. 2012 (Nature) for full details.

Args: interval (GenomicInterval): The interval that you want the FOS for

Kwargs: bgsize (int): The size of the flanking region to use when calcuating the FOS (default: 35)

Returns: A float with the FOS - returns 10000 if it can't calculate it

__getitem__ (*vals*)
Return a dictionary with the cut counts. Can be used in two different ways:

You can either use a string or a GenomicInterval to query for cuts. Returns reads dict with "+" corresponding to the +ve strand and "-" has the data with the -ve strand (rotated 180 degrees)

Args: vals: either a string with the format "chr18,500:600,+" or a GenomicInterval object

```
>>> BAMHandler(example_reads())["chr6,170863142,170863150,+"]
{'+': array([ 1,  0,  0,  0,  1, 11,  1,  0]), '-': array([0, 1, 0, 0, 1, 0, 0, 1])}
>>> BAMHandler(example_reads())["chr6,170863142,170863150,-"]
{'+': array([1, 0, 0, 1, 0, 0, 1, 0]), '-': array([ 0,  1, 11,  1,  0,  0,  0,  1])}
```

__init__ (*filePath*, *caching=True*, *chunkSize=1000*)
Initializes the BAMHandler with a BAM file

Args: filePath (str): the path of a sorted, indexed BAM file from a DNase-seq experiment

Kwargs: chunkSize (int): and int of the size of the regions to load if caching (default: 1000) caching (bool): enables or disables read caching (default: True)

Raises: IOError

3 Handling Genomic Intervals

3.1 GenomicInterval

The `GenomicInterval` is effectively pyDNase's way of storing a BED interval. There are three mandantory fields when creating a new `GenomicInterval`:

```
>>> import pyDNase
>>> interval = pyDNase.GenomicInterval("chr1",100,200)
>>> print interval
chr1        100     200     Unnamed1        0.0     +
```

class `pyDNase.`**GenomicInterval** (*chrom*, *start*, *stop*, *label=0*, *score=0*, *strand='+'*)
Basic Object which describes reads region of the genome

__init__ (*chrom*, *start*, *stop*, *label=0*, *score=0*, *strand='+'*)
Initialization routine

Args: chrom (str): the chromosome

start (int): the start of the interval

stop (int): the end of the interval

Kwargs: label: The name of the interval (will be given an automatic name if none entered)

score (float): the score of the interval (default: 0)

strand (str): the strand the interval is on (default: "+")

You might be wondering why this by itself is helpful. It isn't, until you consider that you can use collections of multiple `GenomicInterval` instances in a `GenomicIntervalSet`

3.2 GenomicIntervalSet

Often, one may be interested in querying cut information for large numbers of regions in the genome (all the DHSs, for example). We provide a basic way to organise BED files using a `GenomicIntervalSet` object.

```
>>> import pyDNase
>>> regions = pyDNase.GenomicIntervalSet("pyDNase/test/data/example.bed")
>>> print len(regions)  # How many regions are in the BED file?
1
>>> print regions
chr6        170863142       170863532       0       0.0     +
```

Iterating/indexing the GenomicIntervalSet object returns GenomicInterval objects, which are sorted by their order of creation (so the order of the BED file if importing a BED file). You can sort by any of the other attributes that the GenomicInterval has, for example, to iterate by score,

```
>>> for i in sorted(regions,key=lambda x: x.score):
        print i
```

The key here, is that as well as querying the BAMHandler for cuts using a string, we can also query using a GenomicInterval object

```
>>> reads = pyDNase.BAMHandler("pyDNase/test/data/example.bam")
>>> reads[regions[0]]                                                    #Note: I've truncated thi.
{'+': array([1,0,0,0,1,11,1,0,0,0,0,0,0,1,0,1,1,0,0,0,0,0,2, ...]),
 '-': array([0,1,0,0,1,0 ,0,1,0,0,1,0,0,0,0,0,0,1,0,0,0,0,5,0,0, ...])}
```

For example, one could use this to efficiently calculate the total number of cuts in a DNase-seq dataset using the intervals in a BED file

```
>>> readcount = 0
>>> for interval in regions:
        readcount += reads[interval]["+"].sum() + reads[interval]["-"].sum()
>>> print readcount
3119
```

We have overloaded the + operator you can directly add other `GenomicIntervalSet` or `GenomicInterval` objects, and you can delete intervals using the `del` keyword thus:

```
>>> print regions
chr6        170863142       170863532       0       0.0     +

>>> regions += pyDNase.GenomicInterval("chr10","100000000","200000000", "0", 10, "-")
>>> print regions
chr6        170863142       170863532       0       0.0     +
chr10       100000000       200000000       0       10.0    -

>>> del regions[0]
>>> print regions
chr10       100000000       200000000       0       10.0    -
```

class `pyDNase.`**`GenomicIntervalSet`**(*filename=None*)

Container class which stores and allow manipulations of large numbers of GenomicInterval objects. Essentially a way of storing and sorting BED files.

`__init__`(*filename=None*)

Inits GenomicIntervalSet. You can also specify a BED file path to load the intervals from

Kwargs: filename (str): the path to a BED file to initialize the intervals with

If no `filename` provided, then the set will be empty

`loadBEDFile`(*filename*)

Adds all the intervals in a BED file to this GenomicIntervalSet. We're quite naughty here and allow some non-standard BED formats (along with the official one):

chrom chromStart chromEnd chrom chromStart chromEnd strand chrom chromStart chromEnd name score strand

Any whitespace (tabs or spaces) will be considered separators, so spaces in names cause a problem!

Note: If you don't supply a strand, we infer that it's +ve.

Args: filename: the path to a BED file to load

Raises: IOError

resizeRegions (*toSize*)
Resized all GenomicIntervals to a specific size

Args: toSize: an int of the size to resize all intervals to

4 Footprinting DNase-seq data

Note: We provide wellington_footprints.py as a script, which will automate footprinting for end users. This below information only necessary if you want to do something fancy. You might want to read the documentation in the source for more information.

We provide a simple interface for footprinting in the `pyDNase.footprinting` module. There are two footprinters, `pyDNase.footprinting.wellington` and `pyDNase.footprinting.wellington1D`, which inherits from wellington and overrides the `calculate` method with a 1D version.

If you want to footprint, we provide an easy method to do so. One can import the Wellington object, and get the Wellington footprints for an interval given the reads from a specific experiment.

```
>>> import pyDNase
>>> import pyDNAse.footprinting as fp
>>> regions = pyDNase.GenomicIntervalSet("pyDNase/test/data/example.bed")
>>> reads = pyDNase.BAMHandler("pyDNase/test/data/example.bam")
>>> footprinter = fp.wellington(regions[0],reads)
>>> footprints = footprinter.footprints(withCutoff=-30)
print footprints
chr6        170863264       170863306       Unnamed4        -150.07397301   +
chr6        170863338       170863383       Unnamed5        -47.9227745068  +
chr6        170863404       170863454       Unnamed6        -164.119817804  +
```

These can easily be written to a BED file, for example by

```
>>> with open("output.bed","w") as bedout:
>>>     bedout.write(str(footprints))
```

If you want, you can also extract the raw footprint score.

```
>>> print footprinter.scores
[  0.00000000e+00   0.00000000e+00   0.00000000e+00   0.00000000e+00
0.00000000e+00   0.00000000e+00   0.00000000e+00   0.00000000e+00
0.00000000e+00   0.00000000e+00   0.00000000e+00   0.00000000e+00 ...
```

So you can write the raw footprinting scores to a WIG file if you want to using something like

```
>>> print "fixedStep\tchrom=" + str(footprinter.interval.chromosome) + "\t start="+ str(footprinte
fixedStep    chrom=chr6     start=170863142    step=1
>>> for i in footprinter.scores:
...     print i
0.0
0.0
0.0
```

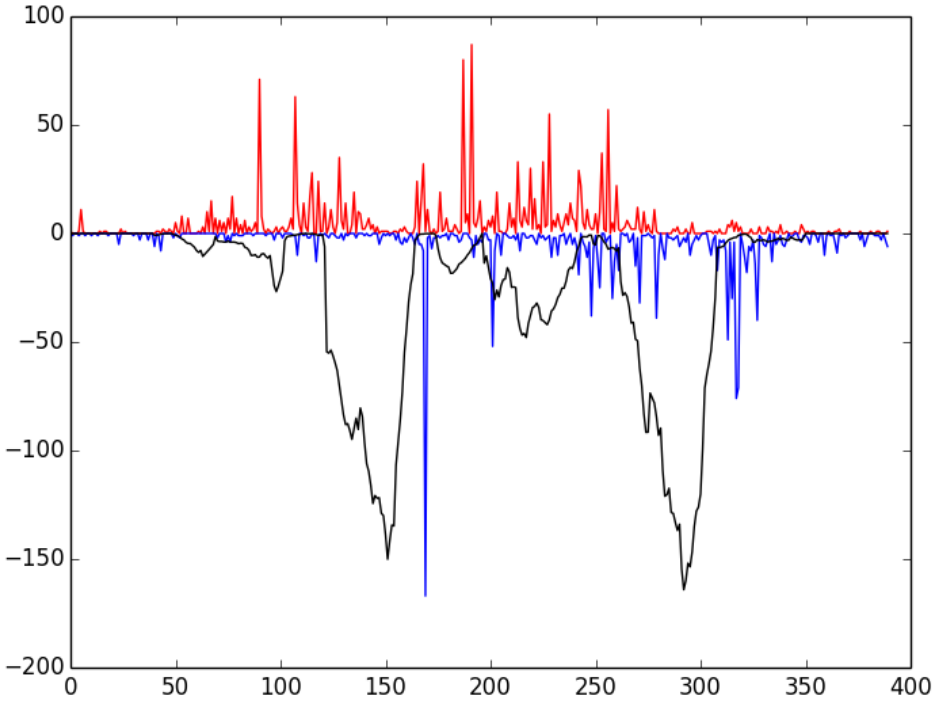
(you'd need to redirect these `print` statements to a file object to write the actualy WIG file)

5 Scripts

pyDNase installs several scripts which also serve as examples of how to use the pyDNAse API. Please have a rummage through the source - it's all documented (and hopefully understandable!)

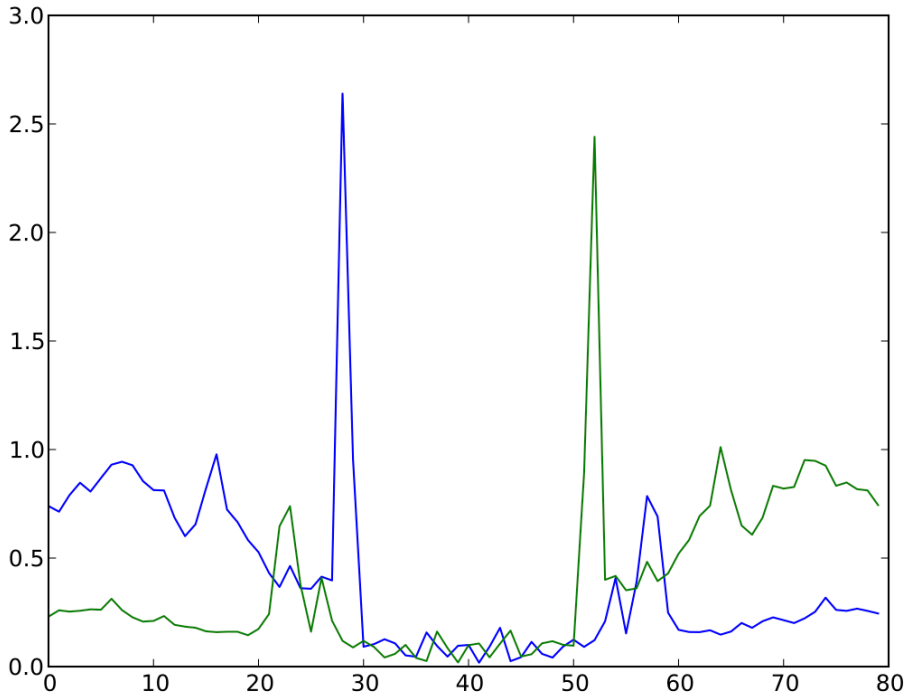
5.1 example_footprint_scores.py

This script tests that everything has been installed and will run correctly. Upon running it, you should see the following window



If so, congratulations! Everything has installed properly. The red and blue bars correspond to cuts on the positive and negative strand, respectively, and the black line represents the raw Wellington footprint scores.

5.2 dnase_average_profile.py



Average profile of DNase I activity surrounding ChIP-seq confirmed CTCF sites in K562 data.

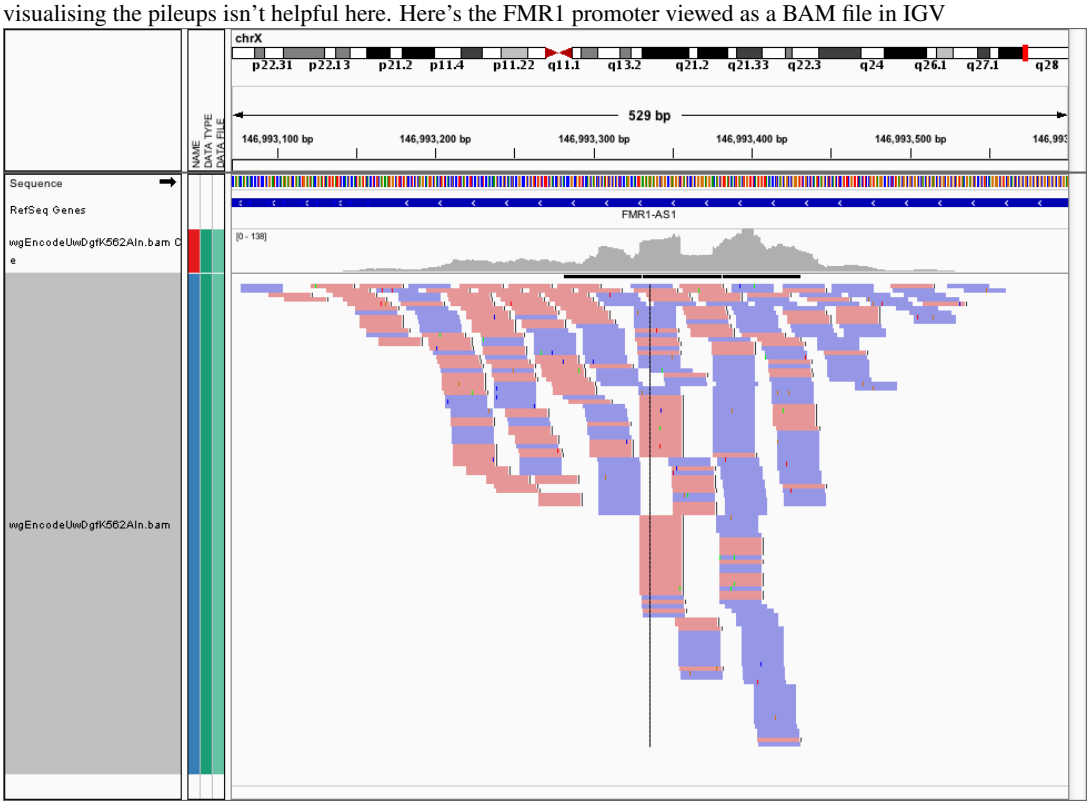
Average profile plots illustrating DNase activity surrounding a set of regions are frequently used in papers. Here, we provide a simple way to generate one

```
usage: dnase_average_profile.py [-h] [-w WINDOW_SIZE] [-i]
                                regions reads output

Plots average profile of DNase activity surrounding a list of regions in a BED
file

positional arguments:
  regions               BED file of the regions you want to generate the
                        average profile for
  reads                 The BAM file containing the DNase-seq data
  output                filename to write the output to

optional arguments:
  -h, --help            show this help message and exit
  -w WINDOW_SIZE, --window_size WINDOW_SIZE
                        Size of flanking area around centre of the regions to
                        plot (default: 200)
  -i                    Ignores any strand information in BED file and plots
                        data relative to reference strand
```

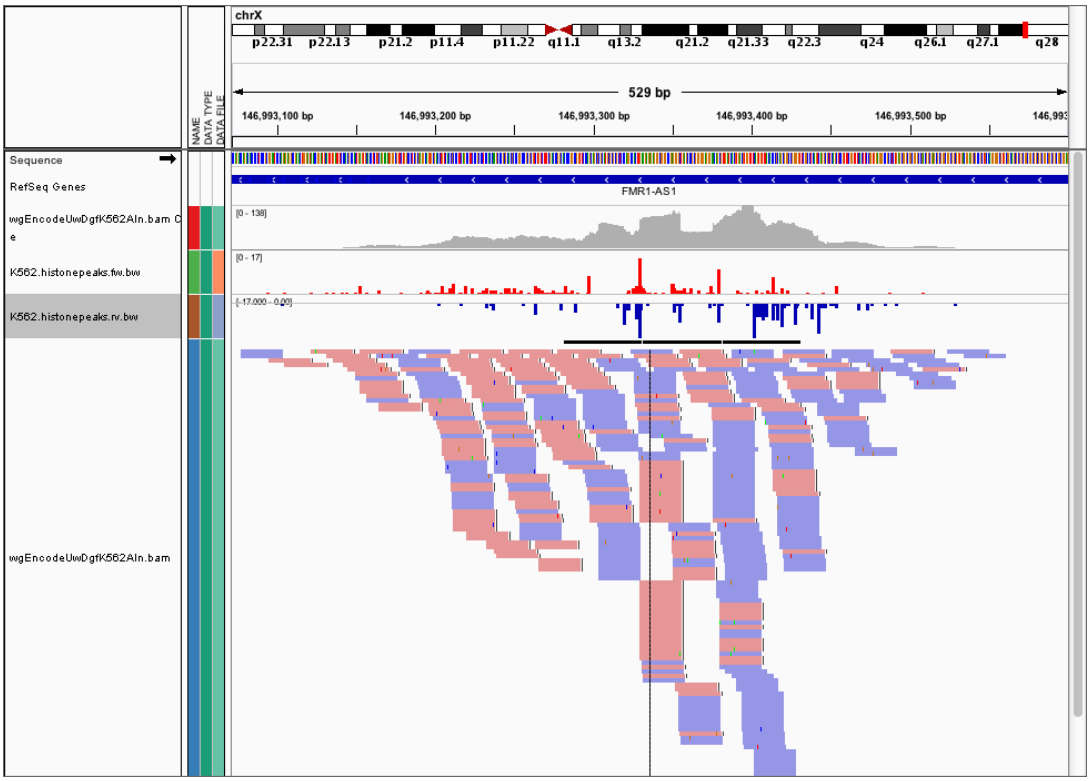
Hopefully this is self-explanatory. This script uses matplotlib to generate the output, so it will write a filetype based on the file extension provided (e.g. `out.png` or `output.pdf`).

5.3 dnase_wig_tracks.py

Often, we want to visualise the raw cut data (just the 5' most ends of the cuts) from a DNase-seq experiment, as visualising the pileups isn't helpful here. Here's the FMR1 promoter viewed as a BAM file in IGV



and here's the corresponding cut locations.

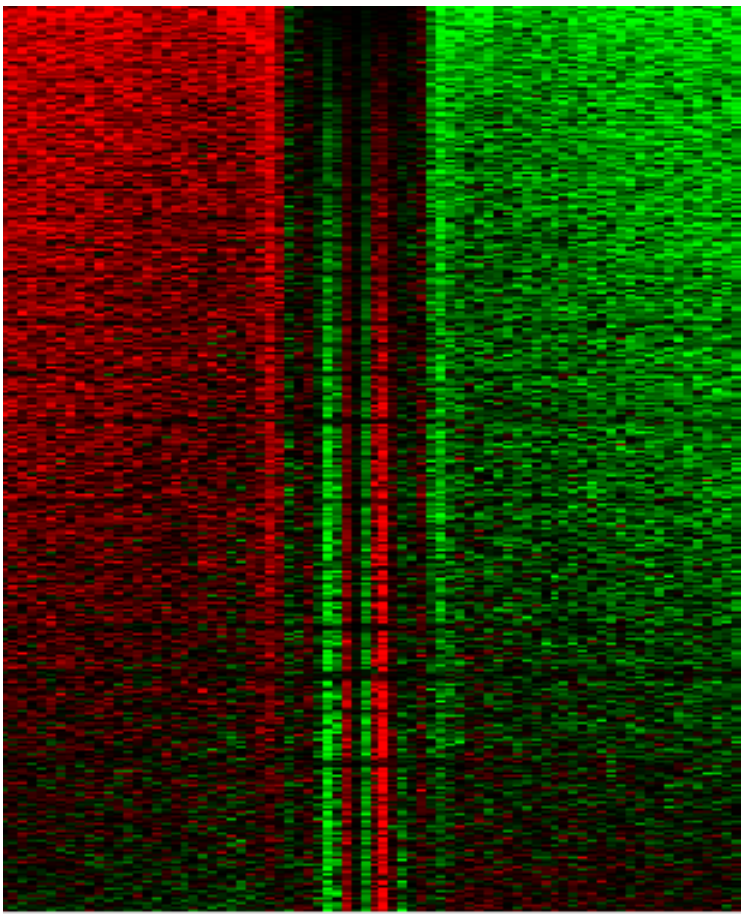


We provide `dnase_wig_tracks.py` that generates a WIG file (we recommend you convert it to a BigWIG file using UCSC's *wigToBigWig*) based on a BAM file a list of regions of interest

```
usage: dnase_wig_tracks.py [-h] [-r] regions reads fw_output rev_output

Writes two WIG files with the cut information based on the regions in reads
BED file and the reads in reads BAM file

positional arguments:
  regions     BED file of the regions you want to write wig tracks for
  reads       The BAM file containing the read data
  fw_output   Path to write the forward reads wig track to
  rev_output  Path to write the reverse reads wig track to

optional arguments:
  -h, --help  show this help message and exit
  -r, --real  Report cuts on the negative strand as positive numbers instead
              of negative (default: 100)
```

Note that by default, cuts on the reverse strand will be reported as negative numbers (for visualisation). If you want to be using this data for something else, you can pass the `-r` flag, which will use the real number of cuts.

5.4 dnase_to_javatreeview.py



Want to make a heatmap? Love JavaTreeView? So do we! This script will generate a CSV file that you can put straight into JavaTreeView to visualize your data.

The options to be aware of here are `-i` and `-a`

```
usage: dnase_to_javatreeview.py [-h] [-w WINDOW_SIZE] [-i] [-o] [-a]
                                regions reads output

Writes a JavaTreeView file based on the regions in reads BED file and the
reads in reads BAM file

positional arguments:
  regions               BED file of the regions you want to generate the
                        heatmap for
  reads                 The BAM file containing the read data
  output                filename to write the CSV output to

optional arguments:
  -h, --help            show this help message and exit
  -w WINDOW_SIZE, --window_size WINDOW_SIZE
                        Size of flanking area around centre of the regions to
                        plot (default: 100)
  -i                    Ignores strand information in BED file
  -o                    Orders output the same as the input (default: orders
                        by FOS)
  -a                    Write absolute cut counts instead strand imbalanced
```

```
                        counts
```

5.5 wellington_footprints.py

So you want to get footprints from your data? No problem. We provide a handy script that will do this for you. There's lots of options here, so please read through them carefully. The most basic usage of the script uses the default parameters described in our original paper. If anything goes wrong at any point, then there should be useful error messages telling you exactly what went wrong.

```
usage: wellington_footprints.py [-h] [-b] [-sh SHOULDER_SIZES]
                                [-fp FOOTPRINT_SIZES] [-d] [-fdr FDR_CUTOFF]
                                [-fdriter FDR_ITERATIONS]
                                [-fdrlimit FDR_LIMIT] [-pv PV_CUTOFFS] [-dm]
                                [-o OUTPUT_PREFIX]
                                regions reads outputdir

Footprint the DHSs in a DNase-seq experiment using the Wellington Algorithm.

positional arguments:
  regions               BED file of the regions you want to footprint
  reads                 The BAM file containing the DNase-seq reads
  outputdir             A writeable directory to write the results to

optional arguments:
  -h, --help            show this help message and exit
  -b, --bonferroni      Performs a bonferroni correction (default: False)
  -sh SHOULDER_SIZES, --shoulder-sizes SHOULDER_SIZES
                        Range of shoulder sizes to try in format
                        "from,to,step" (default: 35,36,1)
  -fp FOOTPRINT_SIZES, --footprint-sizes FOOTPRINT_SIZES
                        Range of footprint sizes to try in format
                        "from,to,step" (default: 11,26,2)
  -d, --one-dimension   Use Wellington 1D instead of Wellington (default:
                        False)
  -fdr FDR_CUTOFF, --FDR_cutoff FDR_CUTOFF
                        Write footprints using the FDR selection method at a
                        specific FDR (default: 0.01)
  -fdriter FDR_ITERATIONS, --FDR-iterations FDR_ITERATIONS
                        How many randomisations to use when performing FDR
                        calculations (default: 100)
  -fdrlimit FDR_LIMIT, --FDR-limit FDR_LIMIT
                        Minimum p-value to be considered significant for FDR
                        calculation (default: -20)
  -pv PV_CUTOFFS, --pv_cutoffs PV_CUTOFFS
                        Select footprints using a range of pvalue cutoffs
                        (default: -10,-20,-30,-40,-50,-75,-100,-300,-500,-700
  -dm, --dont-merge-footprints
                        Disables merging of overlapping footprints (Default:
                        False)
  -o OUTPUT_PREFIX, --output_prefix OUTPUT_PREFIX
                        The prefix for results files (default:
                        <reads.regions>)
```

6 Frequently Asked Questions

Here are common questions we get. If there are any questions about pyDNase or general DNase-seq analysis, either raise an issue on GitHub or email me on `j.piper@warwick.ac.uk`

6.1 How can I identify hypersensitive sites in DNase-seq data?

To identify DNase I hypersensitive sites in DNase-seq data, we recommend using HOMER's `findPeaks` with the parameters: `findPeaks -region -size 500 -minDist 50 -o auto -tbp 0`, converting the HOMER peaks to a BED file using `pos2bed.pl` and then merging the overlapping regions with:

```
$ bedtools sort -i <input.bed> | bedtools merge -i > <output.bed>
```

We find the results are almost exactly the same as the HOTSPOT method employed by ENCODE. See the HOMER documentation for detailed information on how to carry out this procedure.

Nucleic Acids Research, 2014 **1**
doi: 10.1093/nar/gku727

Corrigendum

Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data

Jason Piper, Markus C. Elze, Pierre Cauchy, Peter N. Cockerill, Constanze Bonifer and Sascha Ott

Nucl. Acids Res. (2013) 41 (21): e201. doi: 10.1093/nar/gkt850

It has come to the authors' attention that several presentation errors exist in this article that incorrectly describe the Wellington method.

The formula on page 3 that reads

$$\text{p-value} = \{1 - F[FP^+, FP^+ + SH^+, l_{FP}/(l_{FP} + l_{SH})]\} \cdot \{1 - F[FP^-, FP^- + SH^-, l_{FP}/(l_{FP} + l_{SH})]\}$$

should read as follows

$$\text{p-value} = \{F[FP^+, FP^+ + SH^+, l_{FP}/(l_{FP} + l_{SH})]\} \cdot \{F[FP^-, FP^- + SH^-, l_{FP}/(l_{FP} + l_{SH})]\}$$

We also present the following update to Figure 1, as the original did not specify a logarithmic y-axis, and to clarify the figure legend.

These errors are purely typographical and do not affect the underlying methods, results or conclusions of the manuscript. We apologise for any confusion caused.

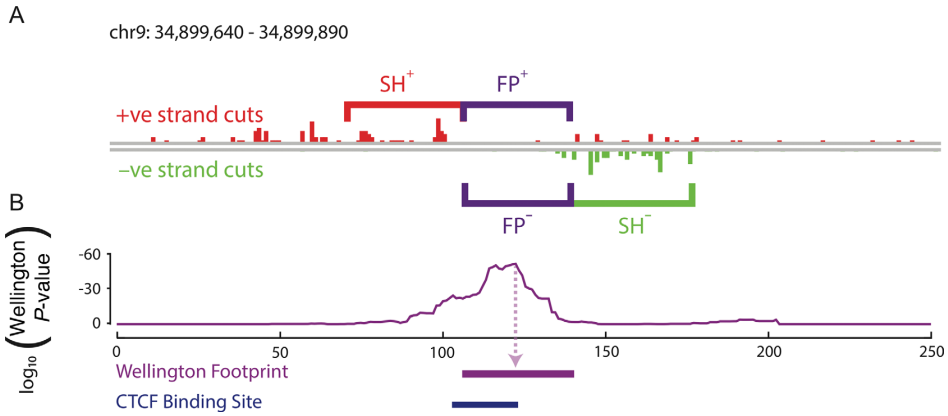


Figure 1. Wellington: a novel strand sensitive algorithm for the identification of protein–DNA binding sites from DNase-seq data. (A) The Wellington algorithm calculates *P*-values for every base pair in all DNase hypersensitive sites in a given DNase-seq data set, where the *P*-value is assigned to the base pair at the centre of the footprint. For each base pair, Wellington tests the hypothesis that there are significantly fewer reads aligning to the forward reference strand footprint region (FP^+) than to the forward reference strand in the upstream shoulder region (SH^+) and significantly fewer reads aligning to the reverse reference strand footprint region (FP^-) than to the reverse reference strand in the downstream shoulder region (SH^-). (B) Example output of the Wellington algorithm. The corresponding footprint prediction recapitulates the ChIP-seq confirmed CTCF-binding site.

Chapter 3

Differential DNase-seq footprinting

3.1 Motivation

After the development of Wellington [2], it was realised that Wellington and pyDNase could be expanded in order to interrogate the data in a more complex fashion. A natural extension to footprinting is the ability to identify footprints that differ between two datasets (termed 'differential' footprints). Whilst it is possible to identify differential footprints using the Wellington algorithm on two datasets and then identifying the complement between the sets of footprints, this approach does not provide a similarity metric of footprint structure between two datasets.

Here, the work performed in Chapter 2 is continued by developing an extension to Wellington called Wellington-bootstrap. This extension of the method scores footprints based on their similarity between two DNase-seq datasets and can be used to identify differential footprints. The power of this analysis is demonstrated by the ability for the differential footprints identified in gene promoters to be attributes to changes in gene expression independent of changes in overall levels of DNase sensitivity. Multiple DNase-seq datasets generated from clinical samples by the NIH Roadmap Epigenomics project were analysed using Wellington-bootstrap. An important result from this study is the finding that the analysis of motif enrichment in differential footprints reveals transcriptional regulators known to be key drivers of cellular identity in these cell populations.

Improvements to pyDNase were also introduced in this paper: pyDNase was extensively benchmarked and portions of code were rewritten in C and the overall method reimplemented to take advantage of parallel computation. This had

a dramatic effect on computation time, as the computation time of footprinting studies scales quasi-linearly with the number of processors. With the recent availability of DNase-seq data on 'naked' genomic DNA, correction for the DNase-seq cutting bias when visualising DNase-seq data was also introduced, providing evidence against the theory that transcription factor binding induces hypersensitivity at specific base pairs due to conformational changes to the DNA [61, 66]. In addition, further documentation was added to the pyDNase-seq library in the form of a 'footprinting tutorial.'

3.2 Contributions

For this paper, I designed the study, developed and implemented the Wellington-bootstrap method, performed the DNase-seq and RNA-seq data analyses, wrote the manuscript, and generated all the figures apart from Figure 2. Salam Assi and Pierre Cauchy designed and implemented the motif clustering analysis used in Figure 2, and also designed the figure. Christophe Ladroue contributed Supplementary Figure 4. Pierre Cauchy, Constanze Bonifer, and Peter Cockerill assisted with the biological interpretation of the differential footprinting results. All authors contributed towards the preparation of the manuscript.

Differential DNase-seq footprinting identifies cell-type determining transcription factors

Jason Piper[1,2], Salam A. Assi[2], Pierre Cauchy[2], Christophe Ladroue[3], Peter N. Cockerill[4,*], Constanze Bonifer[2,*], Sascha Ott[1,*]

1. Warwick Systems Biology Centre, University of Warwick, Coventry, CV4 7AL, UK
2. School of Cancer Sciences, Institute of Biomedical Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, B15 2TT, UK
3. Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK
4. School of Immunity and Infection, Institute of Biomedical Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, B15 2TT, UK

* To whom correspondence should be addressed

submitted to Genome Research

Abstract
The analysis of differential gene expression is a fundamental tool to relate gene regulation with specific biological processes. In order to link expression changes with changes in transcription factor (TF) binding we introduce the concept of differential footprinting alongside a computational tool. We demonstrate that differential footprinting is associated with differential gene expression and can be used to define cell types by their specific TF occupancy patterns.

Main text
Digital DNaseI footprinting is a high throughput adaptation of classical DNaseI footprinting[1]. By subjecting nuclei to digestion by DNaseI, nucleosome-depleted genomic regions (accessible chromatin) that are sensitive to cleavage can be identified as DNase Hypersensitive Sites (DHSs)[2,3]. Analyses of the patterns by which DNase I cuts within DHSs enables the identification of regions protected from digestion or "footprints", which accurately demarcate transcription factor binding sites (TFBSs) at sub-30bp resolution[4,5,6,7,8,9,10,11]. However, all currently available footprinting tools are designed for the analysis of a single DNase-seq data set at a time and thus will indiscriminately identify TFBSs that are part of a variety of different gene regulatory networks, limiting the ability to link regulatory events to cell- and tissue-specific processes, such as changes in cell fate or response to extracellular signals. For gene expression studies, a plethora of computational methods have been developed in order to identify genes that are differentially expressed in different conditions, thereby linking gene expression to changes in cellular status. However, a similar methodology that identifies differential transcription factor occupancy has so far been lacking. Here we describe the development of a novel computational tool to identify differential footprints (DFPs). We show that this tool can be used to link differential TF occupancy with differential gene expression and to identify closely related cell types by virtue of their TF occupancy patterns.
Building on the Wellington footprinting method for single data sets[8], we have developed a conceptually simple, computationally efficient extension, *Wellington-bootstrap*, for pairwise analysis of DNase-seq data sets. Briefly, footprints in data set *A* are detected and at each footprint locus a statistical test is performed testing whether pooling the data of data set *B* with *A* contributes to the footprint pattern or not. This yields a set of sites that are over-footprinted in *A* (under-footprinted in *B*) and associated DFP scores. Repeating the analysis with reversed roles for *A* and *B* yields over-footprinted sites in *B* (under-footprinted in *A*). We chose the approach of pooling data at individual loci in order to avoid biases that may be brought about by variations in sequencing depth.
Applying Wellington-bootstrap to publically available DNase-seq data for CD8+ and CD19+ cells we find 37,488 sites with evidence for DFPs. Furthermore, the Wellington-bootstrap score provides a way to order DFPs by the extent of

footprint differences (**Supplementary Fig. 1**). We found similar results making pairwise comparisons for all DNase-seq data sets for seven cell types from clinical tissue samples. A large proportion (up to 98.5 percent, 43.9 percent on average) of DFPs are found in DHSs that are shared between cell types, in particular in closely related cell types, indicating that these differences would be missed by restricting analyses to the presence or absence of DHSs (**Supplementary Table 1**).

Using spine and CD4+ cells as example we tested the ability of DFPs to re-discover known regulatory links and predict gene expression. In CD4+ cells, the T cell specific TF T-bet binds T-box motifs and enhances target gene expression as part of the Th1-differentiation programme[12]. In spine, the TF MAZ is known to be involved in neuronal development[13]. Among the set of all DFPs located near transcriptional start sites and over-footprinted in CD4+ cells we identified the sites containing a match for the T-box motif. We found that the expression of nearby genes differed significantly, with the DNase-seq data providing strong evidence for the presence of protein binding in CD4+ cells and absence of binding in spine (**Fig. 1a,b**). Similarly, we found that a link between binding to MAZ motifs and gene expression was evident (**Supplementary Fig. 2a,b**), demonstrating the ability of the DFP approach to isolate the effect of individual TFs from their genomic context.

Previously, comparisons of total read numbers in DHSs have been used as a means of analysing pairs of DNase-seq data sets[14]. We identified the set of T-box motif-containing DHSs in gene promoters with the highest increase in read numbers in CD4+ cells compared to spine. While these showed differential expression of nearby genes, no evidence for differences in binding was revealed using this approach (**Fig. 1c,d**). Similarly, this approach did not reveal the regulatory link between MAZ binding and target gene expression (**Supplementary Fig. 2c,d**). The cleavage profiles shown in **Fig. 1b,d** and **Supplementary Fig. 2b,d** have been corrected for the known sequence preference of the DNaseI enzyme. **Supplementary Fig. 3** compares cleavage profiles with and without this correction. Overall, this suggests that unlike DFPs, motif analysis of DHSs is insufficient to link a given TF to changes in gene expression, making the use of DFPs a valuable tool for this purpose.

We sought to further explore the potential of the DFP approach to reveal cell type-specific regulatory mechanisms. Using differential footprints amongst all pairs of DNase-seq data sets of seven primary cell types, we determined the relative frequency of motif occurrences for a set of known TF binding motifs and used this data to cluster the set of pairs of cell lines as well as the set of TF binding motifs (**Fig. 2**). This analysis generated a number of striking results. Firstly, our DFP methodology combined with clustering recovered the different cell types as they formed separate clusters. Moreover, it was able to distinguish related cell types such as CD19+ B cells, T cells and CD14+ monocytic cells all of which belong to the hematopoietic lineage. In addition, this analysis was capable

of differentiating between sub-types of cells of the same lineage, such as CD4+ helper and CD8+ cytotoxic T cells. Secondly, the analysis gave interesting insights into the relative role of individual TF families within a given cell type. For example, high differential C/EBP motif occupancy was a classifier for CD14+ monocytes as well as fibroblasts, both of which express C/EBPa, but the relative motif frequency was lower in fibroblasts which agrees with the fact that this factor is absolutely essential for monocyte but not fibroblast development[15,16]. Another interesting finding was that increased occupancy of PU.1 motifs was a classifier for both B cells and CD14+ monocytic cells where this factor plays an important role[17], but a significant number of sites were occupied also in T cells. PU.1 is not expressed in T cells and its overexpression is detrimental for their development[18] indicating that PU.1 motifs originally bound in hematopoietic stem cells are now occupied by a different factor of the Ets family.

To facilitate the wide-spread use of our method, we provide an implementation of Wellington-bootstrap alongside a substantial update of pyDNase, including increased performance and parallelised computations. This is released as open source under the GPLv3 license at http://jpiper.github.io/pyDNase/ (upon publication).

In conclusion, we introduce a fundamental and useful method for differential footprints, provide a tool for the detection of DFPs, and reveal the potential of this approach to map regulators to context-specific gene expression. Applying this methodology will be highly relevant for classifying closely related cell types, both in the normal, but also the diseased state and to assess the relative importance of specific TF families for each state. Wellington-bootstrap is applicable to any pair of DNase-seq data sets obtained with comparable experimental protocols including perturbation and time course experiments, making it a widely applicable approach for the identification of transcriptional regulatory hierarchies.

Acknowledgments

J.P. was in part supported by the Engineering and Physical Sciences Research Council (EP/P50578X/1 PhD grant). Work in the labs of C.B. and P.N.C. was supported by grants from Leukaemia Lymphoma Research and the Kay Kendall Leukaemia Fund. We thank Alan Boyle, Shirley Liu and Cliff Meyer for stimulating discussions.

Author contributions

J.P. designed the study together with S.O. and S.A.A.. J.P., S.A.A., P.C., and C.L. performed data analyses. J.P. developed the software and wrote the manuscript. All other authors contributed towards the drafting of the manuscript. P.N.C., C.B., and S.O. provided general guidance and supervision.

Competing financial interests

The authors declare no competing financial interests.

References

[1] **Galas *et al.*, *Nucleic Acids Res* 5, 3157-3170 (1978).**
[2] **Boyle *et al.*, *Cell* 132, 311-322 (2008).**
[3] **Cockerill, *FEBS J* 278, 2182-2210 (2011).**
[4] **Hesselberth *et al.*, *Nat Methods* 6, 283-289 (2009).**
[5] **Pique-Regi *et al.*, *Genome Res* 21, 447-455 (2011).**
[6] **Boyle *et al.*, *Genome Res* 21, 456-464 (2011).**
[7] **Neph *et al.*, *Nature* 489, 83-90 (2012).**
[8] **Piper *et al.*, *Nucleic Acids Res* 41, e201 (2013).**
[9] **Sherwood *et al.*, *Nat Biotechnol* 32, 171-178 (2014).**
[10] **Sung *et al.*, *Mol Cell* 56, 275-285 (2014).**
[11] **He *et al.*, *Nat Methods* 11, 73-78 (2014).**
[12] **Szabo *et al.*, *Cell* 100, 655-669 (2000).**
[13] **Wang *et al.*, *Mol Neurobiol* 47, 228-240 (2013).**
[14] **He *et al.*, *Genome Res* 22, 1015-1025 (2012).**
[15] **Zhang *et al.*, *Proc Natl Acad Sci U S A* 94, 569-574 (1997).**
[16] **Ranjan *et al.*, *Oncogene* 28, 3235-3245 (2009).**
[17] **Scott *et al.*, *Science* 265, 1573-1577 (1994).**
[18] **Anderson *et al.*, *Immunity* 16, 285-296 (2002).**

Cell type A	Cell type B	DHSs in A	DHSs in B	DHSs shared between A and B	Sites over-footprinted in A	Sites in common DHSs over-footprinted in A	Sites over-footprinted in B	Sites in common DHSs over-footprinted in B
CD4	CD8	84,830	60,890	49,365	14,772	10,600 (71.8)	3,874	3,584 (92.5)
CD4	CD14	84,830	109,647	47,887	14,819	6,219 (42)	17,932	7,663 (42.7)
CD4	CD19	84,830	89,660	43,282	18,525	10,423 (56.3)	19,439	13,018 (67)
CD4	CD56	84,830	69,966	54,739	17,745	14,611 (82.3)	2,616	2,526 (96.6)
CD4	Spine	84,830	197,751	34,812	24,652	9,158 (37.1)	93,152	10,233 (11)
CD4	Fibroblasts	84,830	193,546	40,240	21,473	7,087 (33)	118,265	11,741 (9.9)
CD8	CD14	60,890	109,647	32,185	11,602	6,529 (56.3)	55,650	12,546 (22.5)
CD8	CD19	60,890	89,660	32,350	8,780	5,520 (62.9)	28,708	15,549 (54.2)
CD8	CD56	60,890	69,966	51,965	1,458	1,428 (97.9)	335	330 (98.5)
CD8	Spine	60,890	197,751	27,631	13,128	5,444 (41.5)	110,950	11,330 (10.2)
CD8	Fibroblasts	60,890	193,546	30,237	13,734	5,894 (42.9)	156,418	15,573 (10)
CD14	CD19	109,647	89,660	36,349	48,031	15,909 (33.1)	27,111	18,140 (66.9)
CD14	CD56	109,647	69,966	33,900	54,850	17,845 (32.5)	7,842	5,357 (68.3)
CD14	Spine	109,647	197,751	33,141	53,731	13,584 (25.3)	96,856	13,563 (14)
CD14	Fibroblasts	109,647	193,546	45,179	37,641	8,383 (22.3)	108,482	12,677 (11.7)
CD19	CD56	89,660	69,966	35,766	31,561	19,315 (61.2)	5,553	4,130 (74.4)
CD19	Spine	89,660	197,751	31,858	28,993	13,118 (45.2)	97,388	14,826 (15.2)
CD19	Fibroblasts	89,660	193,546	30,831	32,531	13,760 (42.3)	138,301	20,224 (14.6)
CD56	Spine	69,966	197,751	28,731	8,633	4,404 (51)	110,996	13,892 (12.5)
CD56	Fibroblasts	69,966	193,546	31,469	9,237	4,769 (51.6)	154,923	20,024 (12.9)
Spine	Fibroblasts	197,751	193,546	64,733	24,756	5,497 (22.2)	35,202	9,461 (26.9)

Supplementary Table 1. A large proportion of differential footprints occurs in shared DHSs. Number of DHSs and shared DHSs, number of over-footprinted sites, and number of over-footprinted sites located in the overlap of shared DHSs are shown for pairs of cell types. For closely related cell types most differential footprints tend to be found in common DHSs (e.g. CD4+ vs. CD56+). Developmentally distant cell types, however, often have a large number of DHSs that are cell type specific, and therefore the majority of differential footprints are in cell-type specific DHSs (e.g. CD56+ cells vs. fibroblasts).

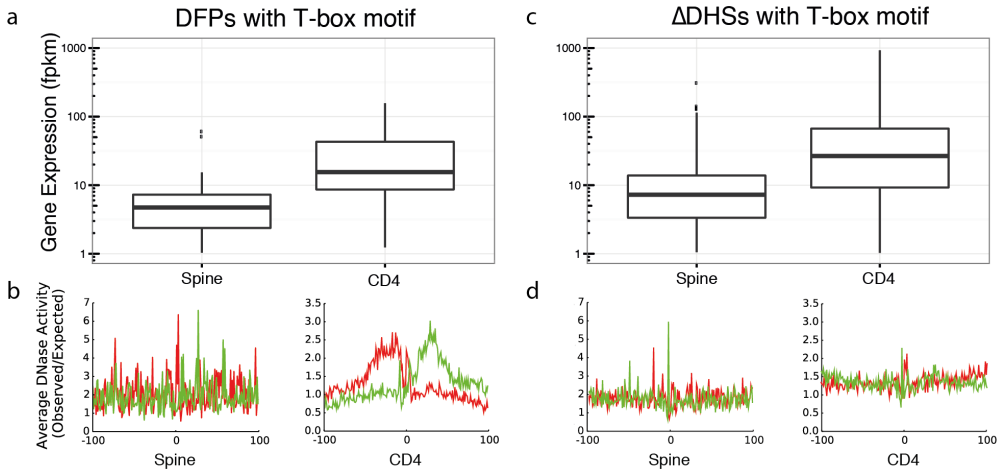


Figure 1. Differential footprints reveal links between TF binding and gene expression. (**a**) Differential gene expression ($p<0.005$, Mann-Whitney U test) of all genes that have a differential CD4 footprint containing a match for the T-box motif in their promoter. (**b**) Average bias-corrected DNase-seq cleavage profiles (red: positive strand reads, green: negative strand reads) for T-box sites in promoters of genes from (**a**) show evidence for binding of T-box motifs in CD4+ cells, but not in spine. Genes over-footprinted for T-box in CD4+ cells are also over-expressed, confirming a known lineage-determining link. (**c**) Differential gene expression of all genes that have a differential CD4+ DHS containing a match for the T-box motif in their promoter. (**d**) Average bias-corrected DNase-seq cleavage profiles for T-box sites in promoters of genes from (**c**) do not show evidence for binding in either cell type. The differential expression observed in (**c**) cannot be linked to TF binding using differential DHS scores alone.

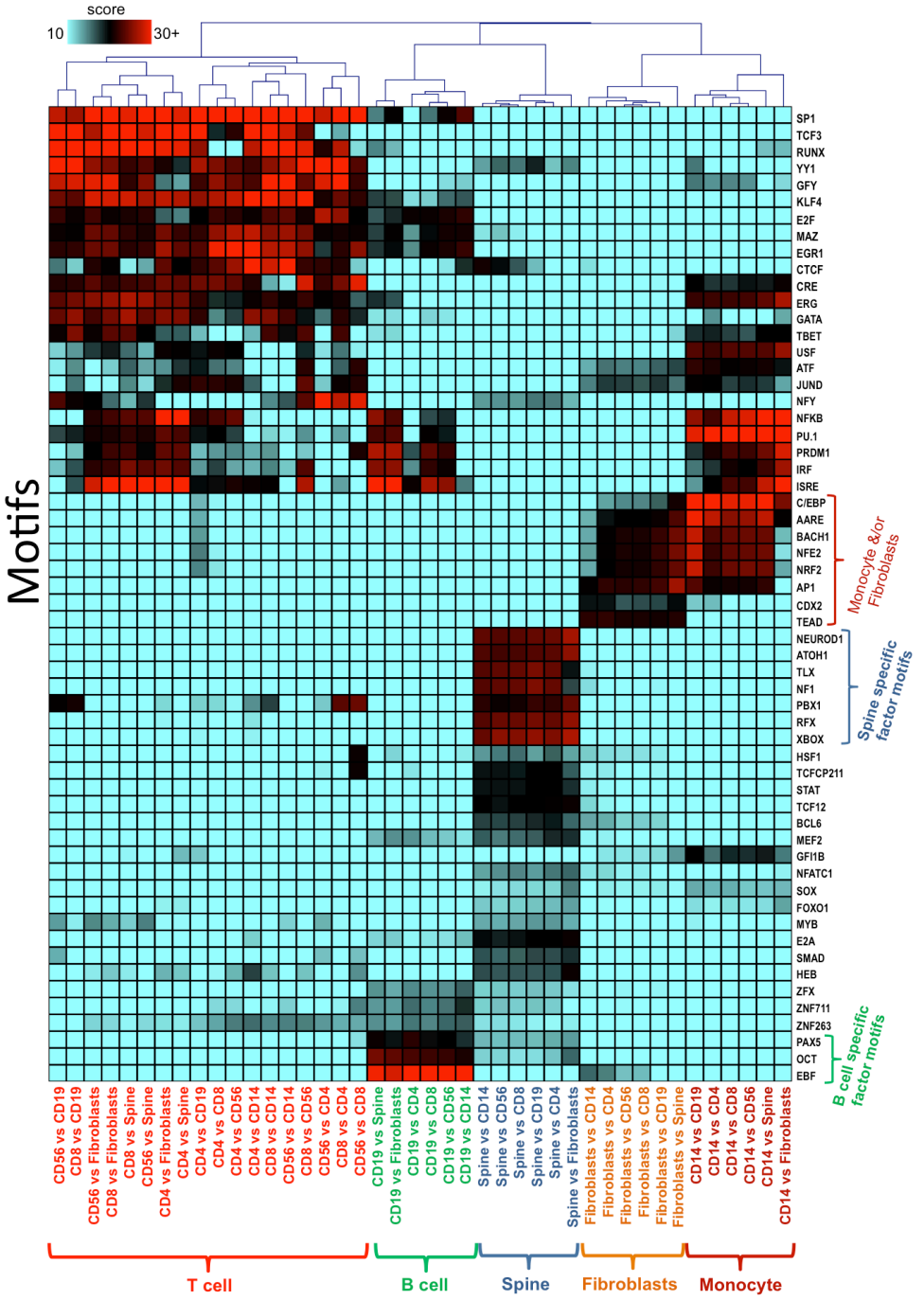
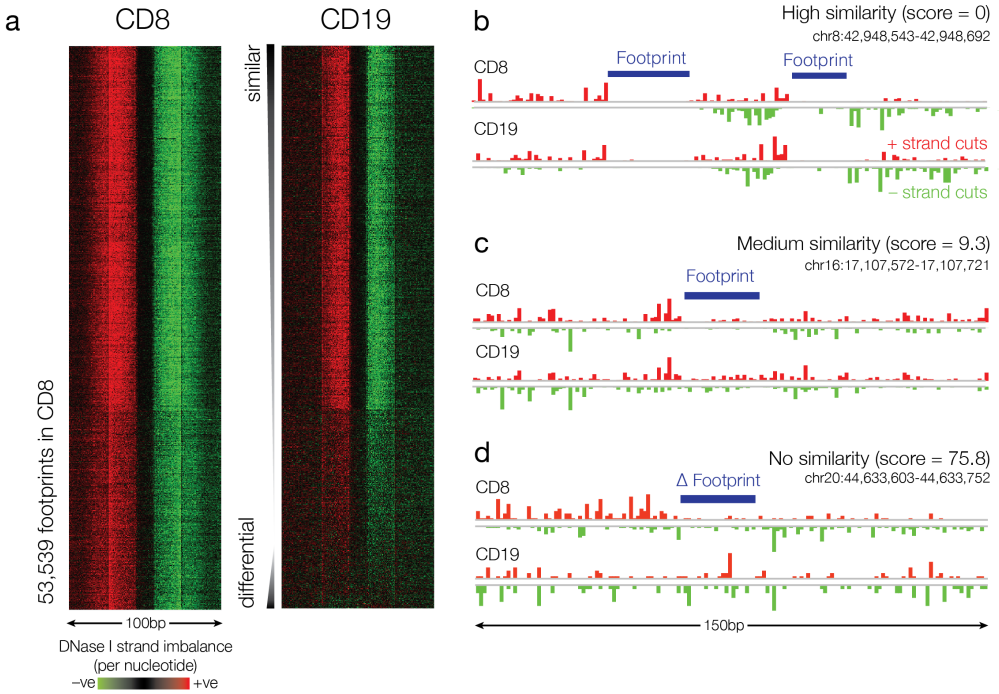


Figure 2. Analysis of differential footprints in the haematopoietic system reveals cell-type specific transcription factor networks. Differential footprints in 42 pairs of cell types and matches to known motifs inside differential footprints were determined using DNase-seq data from the NIH Roadmap Epigenomics project. Coloured boxes represent motif frequency with red indicating higher than average frequency. Hierarchical clustering was applied to rows and columns. The result correctly groups cell types and reveals known and likely regulatory factors.

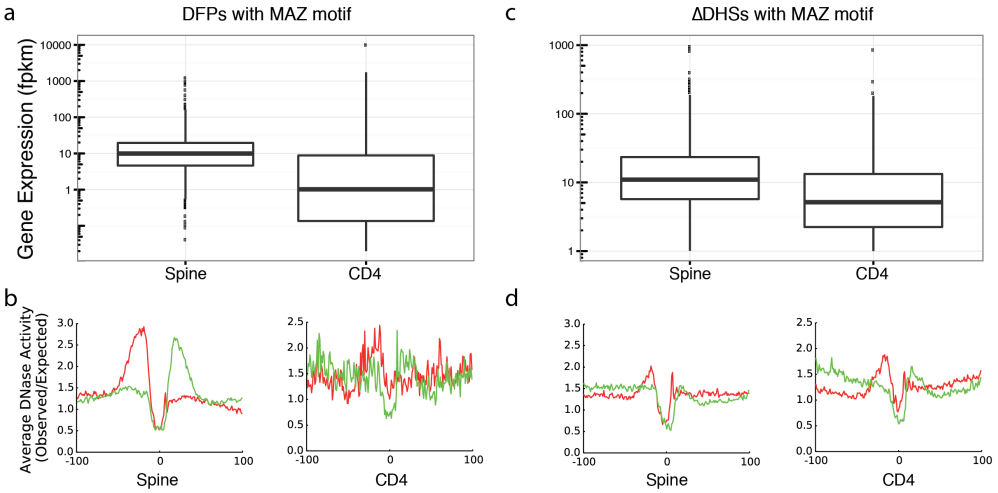
SUPPLEMENTAL FIGURES

Differential DNase-seq footprinting identifies cell-type determining transcription factors

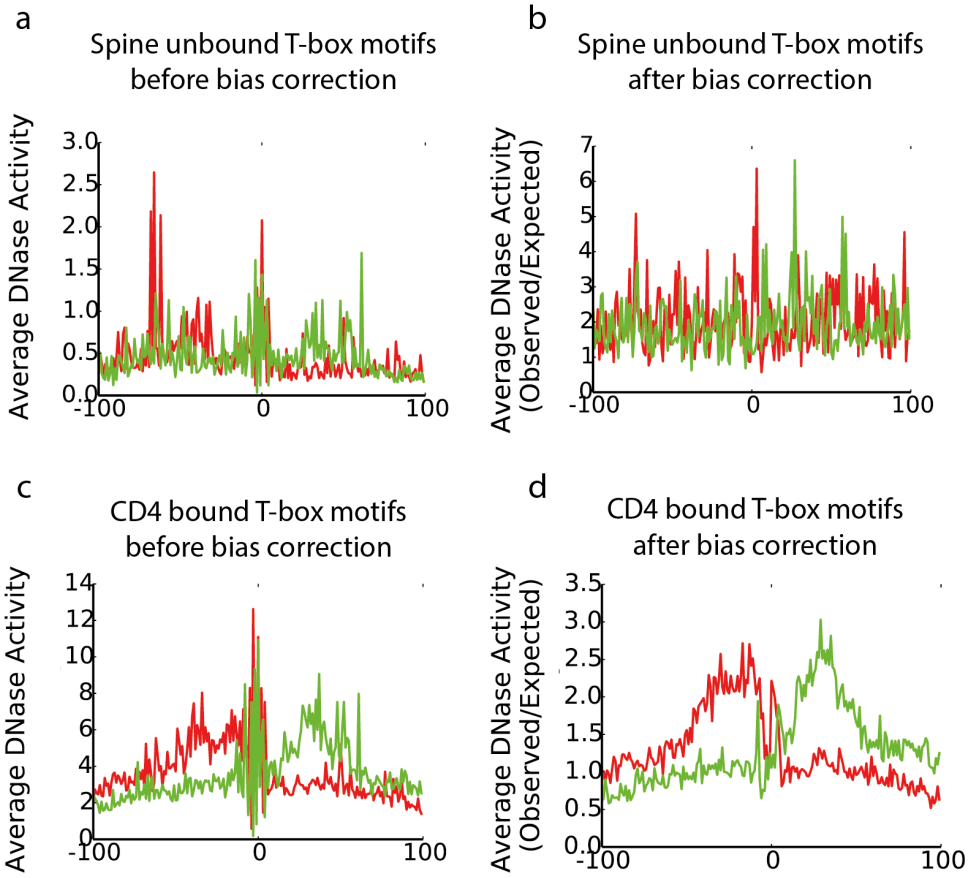
Jason Piper, Salam A. Assi, Pierre Cauchy, Christophe Ladroue, Peter N. Cockerill, Constanze Bonifer, Sascha Ott



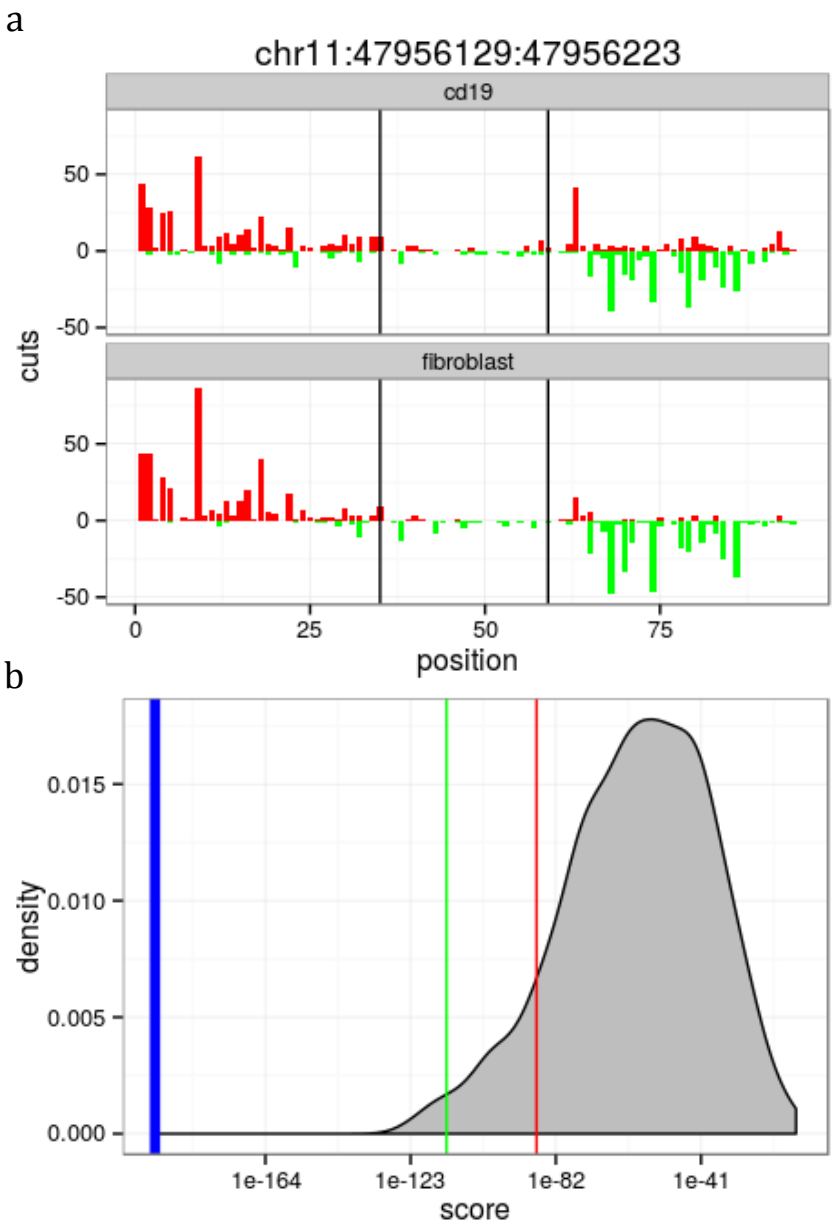
Supplementary Figure 1. Wellington-bootstrap scores differential footprint occupancy between DNase-seq datasets. Wellington-bootstrap was applied at footprint loci in CD8+ cells to detect over-footprinted sites relative to CD19+ cells. (**a**) 53,539 loci were sorted by increasing Wellington-bootstrap score comparing CD8 vs CD19. 8,780 loci were deemed to be DFPs. Red indicates an excess of positive strand cuts over negative strand cuts per nucleotide position, and green indicates an excess of negative strand cuts. Common footprints at the top of the heatmap share similar DNase activity as exemplified in (**b**) whereas footprints with increasing differential score towards the bottom of the heatmap show increasingly differential footprints (**c**,**d**).



Supplementary Figure 2. Differential footprints reveal links between TF binding and gene expression. (**a**) Differential gene expression ($p<0.005$, Mann-Whitney U test) of all genes that have a differential spine footprint containing a match for the MAZ motif in their promoter. (**b**) Average bias-corrected DNase-seq cleavage profiles (red: positive strand reads, green: negative strand reads) for MAZ sites in promoters of genes from (**a**) show evidence for binding of MAZ motifs in spine cells, but not in CD4+ cells. Genes over-footprinted for MAZ in spine cells are also over-expressed, confirming a known lineage-determining link. (**c**) Differential MAZ expression of all genes that have a differential spine DHS containing a match for the MAZ motif in their promoter. (**d**) Average bias-corrected DNase-seq cleavage profiles for MAZ sites in promoters of genes from (**c**) show evidence for binding in both cell types. The differential expression observed in (**c**) cannot be linked to differences in TF binding using differential DHS scores alone.



Supplementary Figure 3. Bias correction refines profiles of average cutting. For T-box-containing loci of differential footprints used in **Figure 1b** average DNaseI cleavage profiles are shown before (**a, c**) and after (**b, d**) correcting for the sequence specificity of DNaseI cleavage using a 6-mer model (He et al., 2013). Plots (**b**) and (**d**) are the ones shown in **Figure 1b**.



Supplementary Figure 4. Example of a footprint deemed non-differential. (**a**) Red (green) bars represent numbers of 5' ends of reads aligning to the positive (negative) reference strand. Vertical black lines indicate footprint region. (**b**) Bootstrap distribution for data shown in (**a**). Nucleotide positions in CD19 data were randomly shuffled and the distribution of Wellington footprint scores after pooling the shuffled CD19 data and the fibroblast data was determined. Blue vertical bar shows the Wellington score after pooling data without shuffling. Green: Wellington footprint score in fibroblast data. Red: footprint score in CD19 data. As pooling without shuffling yields a better footprint score than pooling with shuffling the footprint is considered non-differential.

Differential DNase-seq footprinting identifies cell-type determining transcription factors

Jason Piper, Salam A. Assi, Pierre Cauchy, Christophe Ladroue, Peter N. Cockerill, Constanze Bonifer, Sascha Ott

Online Methods

DNase-seq data and peak-finding

DNase-seq data from the NIH Roadmap Epigenomics project[1] were downloaded from the Short Read Archive (accessions CD4: SRX214041, CD8: SRX204403, CD19: SRX342324, CD14: SRX252602, CD56: SRX204402, spinal column: SRX121287, fibroblasts: SRX135564) and were aligned to hg19 using Bowtie 2.2.0 using the default parameters. DNase hypersensitive site detection for all DNase-seq data was performed using HOMER's findPeaks.pl tool[2] with the parameters "findPeaks -region -size 500 -minDist 50 -o auto -tbp 0".

Differential footprinting – Wellington-bootstrap

Wellington-bootstrap first determines Wellington footprints in the primary dataset. At each footprint locus the data from the comparator dataset is added and the Wellington footprint score for the pooled data evaluated. Wellington-bootstrap then assesses if the change in footprint score is a consequence of the increase in read numbers after pooling reads or if the data from the comparator dataset makes a contribution to the footprint structure. To do this, the comparator data is randomly shuffled 1000 times, pooled, and the Wellington footprint score evaluated (see example in **Supplementary Fig. 4**). Shuffling is done in a strand independent manner, randomising the positions of the counts of 5' DNase cuts per base pair on the positive and negative strand. The score of pooled data without shuffling is assessed against the bootstrap distribution and the percentile used as the differential footprinting score. Low scores indicate non-differential footprints, high scores differential footprints. **Supplementary Fig. 1** shows that sorting by this score orders pairs of footprints in an intuitive manner enabling the user to retrieve the most differential footprints while choosing the stringency. 10 was used as the threshold in this work. The role of the two datasets is reversed and the computation repeated to obtain both over- and under-footprinted sites.

Initially it was thought that a measure of flexibility would be required regarding the width of the footprint and its position in the two datasets. Whilst initial methods were developed to take this into consideration, we found that this provided no improvement to the method, yet yielded a significant speed decrease.

This analysis has been implemented in the wellington_bootstrap.py script as part of pyDNase 0.2.0.

Differential DHSs – Figure 1 and Supplementary Figure 2

Differential DHSs (ΔDHS) scores were calculated according to the method proposed by He et al.[3] and the analysis script has been provided as `dnase_dshs_scores.py` in pyDNase 0.2.0. DHSs were then filtered to those that were within 2kb of a single TSS using the hg19 UCSC knownGene gene model, and the DHSs showing the top and bottom *n*=1000 ΔDHS scores were chosen as the differential DHSs. Equivalent results were obtained using the following alternative choices for *n*: 50 (matching the number of DFPs used in **Figure 1a,b**), top 476 and bottom 300 (corresponding to two standard deviations difference to mean ΔDHS score), 1403 (corresponding to top and bottom 10%).

RNA-seq analysis

RNA-seq data were downloaded from the Short Read Archive (accessions CD4: SRR643766, spinal column: SRR980477) and FPKM was estimated using Tophat 2.0.11 and Cufflinks 2.1.1 with the Illumina iGenomes UCSC hg19 knownGene GTF file.

Motif analysis – Figure 2

The annotatePeaks.pl script of the HOMER[2] package was used to find occurrences of known motifs in peaks. Wellington-bootstrap was applied to compute 42 sets of differential footprints for all ordered pairs of the seven cell types used (CD4/CD8 T-cells, CD56 NK cells, CD19+ B cells, Spine (embryo), fibroblasts, CD14+ monocytes). To analyse motif frequencies in differential footprints motif search was done within the differential footprint coordinates extended by 10bp either side. Relative motif frequencies were calculated as

$$\text{Relative frequency motif } i \text{ in comparison } j = (n_{ij}/M_j)\times\left(C\sum_j M_j/\sum_j n_{ij}\right),$$

where C is a scaling constant, n_{ij} is the number of differential footprints in set j *(j=1,2,...,42)* that are occupied by motif *i (i=1, 2,....,I)*, I is the total number of motifs used, and M_j the total number of differential footprints in each subset j *(j=1,2,...,42).* A matrix was generated and motif scores displayed as a heatmap after hierarchical clustering with Euclidean distance and complete linkage. Blue indicates low relative frequency; red/black indicates high relative frequency. Heatmaps were generated using Mev of the TM4 microarray software suite[4].

pyDNase 0.2.0 – cutting bias correction

In order to plot cut bias corrected average DNase cleavage plots, the DNaseI 6-mer cutting bias data from naked genomic data from the IMR90 cell line and for each region an 'expected count' was calculated using the 'predicted count' formula from He et al. 2013[3]. The observed cuts at each base pairs were then

divided by the expected counts. Bias correction modes have been added to the plotting scripts in pyDNase that can be invoked with the '`-b <genome.fa>`' option. The `BAMHandlerWithBias` class in pyDNase provides underlying access to the bias correction for power users. In this we have provisioned the ability for the user to supply a Variant Call Format (VCF) file so that the reference DNA sequence can be corrected using SNPs present in the sample being analysed if desired.

pyDNase 0.2.0 – other new features and improvements

pyDNase 0.2.0 represents a major release for pyDNase, bringing several improvements. The core Wellington algorithm was reimplemented in C, and the underlying code structure was refactored in order to allow for parallelisation of Wellington score calculation. On a dual 2.66Ghz i7 Xeon workstation with 8 cores, footprinting a single dataset takes approximately 30 minutes, compared to up to 20 hours previously on a single core – this performance increase scales linearly with number of cores utilised. In addition, a number of analysis scripts have been added to the pyDNase library for calculating ΔDHS scores, calculating Wellington-bootstrap scores, annotation of BED files with Footprint Occupancy Scores, and the annotation of a BED file with DNase cuts. A comprehensive DNase-seq footprinting tutorial has also been added to assist those new to DNase-seq analysis and DNase-seq footprinting. Full details can be found at the pyDNase github repository (https://github.com/jpiper/pyDNase/).

References

1 Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-1048, doi:nbt1010-1045 [pii] 10.1038/nbt1010-1045 (2010).

2 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:S1097-2765(10)00366-7 [pii] 10.1016/j.molcel.2010.05.004 (2010).

3 He, H. H. *et al.* Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Res* **22**, 1015-1025, doi:gr.133280.111 [pii] 10.1101/gr.133280.111 (2012).

4 Saeed, A. I. *et al.* TM4 microarray software suite. *Methods Enzymol* **411**, 134-193, doi:S0076-6879(06)11009-5 [pii] 10.1016/S0076-6879(06)11009-5 (2006).

pyDNase 0.2.0 Footprinting Tutorial

Jason Piper

December 06, 2014

Contents

1 DNase I cleavage bias correction **2**
1.1 How is this useful? . . . 3
1.2 Great! What's the catch? . . . 3

2 Frequently Asked Questions **3**
2.1 How can I identify hypersensitive sites in DNase-seq data? . . . 3
2.2 pyDNase won't install/import or gives weird errors . . . 4
2.3 These footprints from are too stringent for my liking . . . 4

3 DNase-seq footprinting Tutorial **4**
3.1 Testing pyDNase installation . . . 4
3.2 Getting data in the correct format . . . 5
3.3 Aligning your reads . . . 5
3.3.1 Dealing with SRA . . . 5
3.3.2 Sorting and Indexing . . . 5
3.4 Peak calling . . . 6
3.5 Quick and easy Footprinting . . . 6
3.6 Interpreting Wellington's Output . . . 7
3.7 Visualising the data . . . 7
3.8 Visualising Footprints as average plots . . . 8
3.9 Visualising Footprints as heat maps . . . 9
3.10 Motif Finding . . . 10
3.11 Leveraging pyDNase for fun and profit (mainly fun) . . . 11

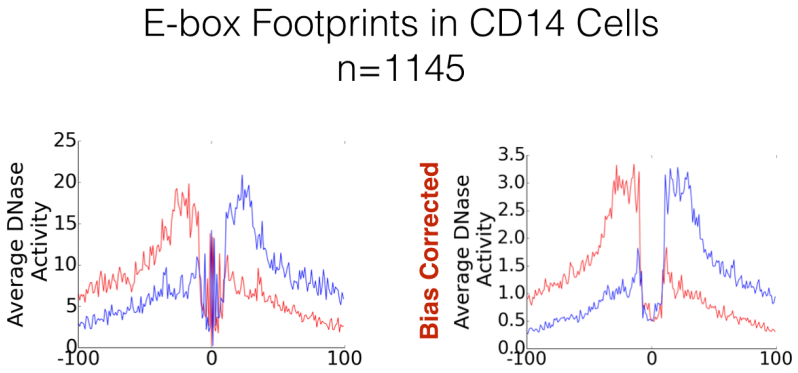
1 DNase I cleavage bias correction

pyDNase 0.2.0 comes with consideration for DNase-seq cutting bias. At the moment, this is a preliminary implementation limited to just the visualisation functions `dnase_average_profile.py` and `dnase_to_javatreeview.py`. Instead of DNase-seq read counts, this instead reports the fold change over a theoretical background model using the bias values reported in the IMR90 naked genomic DNA reported in He et al. 2014

This is currently not a part of the core Wellington footprinting function, as a more thorough analysis of the impact of introducing such a feature on the method is needed here.

1.1 How is this useful?

Often, people use `dnase_average_profile.py` and `dnase_to_javatreeview.py` to visualise the footprints of a specific transcription factor. When this is the case, the plot will be centred on a common motif, and sequence specific cutting biases have the potential to become very apparent. As an example, footprinted E-box motifs have a heightened cutting profile in the centre of the footprint. These were theorised by Neph et al. 2012 to be caused by conformational changed to the DNA induced by the binding of the transcription factor. However, we can show here that this is not that case, and by accounting for sequence bias these patterns dissapear.



1.2 Great! What's the catch?

There are two catches here. The first is that this doesn't work well on regions with low numbers of reads. If your read depth is really low for the regions you are trying to plot, your data might not look good. Second, this is no longer reporting read counts, but a "fold change". This is more of an implementation issue as we felt proper methods for accounting for DNase cleavage bias are best left for another study.

2 Frequently Asked Questions

Here are common questions we get. If there are any questions about pyDNase or general DNase-seq analysis, either raise an issue on GitHub or email me on `j.piper@warwick.ac.uk`

2.1 How can I identify hypersensitive sites in DNase-seq data?

To identify DNase I hypersensitive sites in DNase-seq data, we recommend using HOMER's `findPeaks` with the parameters: `findPeaks -region -size 500 -minDist 50 -o auto -tbp 0`, converting the HOMER peaks to a BED file using `pos2bed.pl` and then merging the overlapping regions with:

```
$ bedtools sort -i <input.bed> | bedtools merge -i - > <output.bed>
```

We find the results are almost exactly the same as the HOTSPOT method employed by ENCODE. See the HOMER documentation for detailed information on how to carry out this procedure.

2.2 pyDNase won't install/import or gives weird errors

The most common issue here is that you have old versions of the dependencies - namely `scipy`, `numpy`, or `pysam` installed - try updating these to their latest version. pyDNase is built against Python 2.6, 2.7 and 3.0 by the Travis contious integration (CI) system, so we're very confident in the deployability of the codebase.

2.3 These footprints from are too stringent for my liking

This is a common question - if you have low read depths you might need to adjust the `-fdrlimit` parameter to something less stringent like `"-10"` or `"-5"`, which sets the mimimum amounts of evidence required to support the alternate hypothesis of there being a footprint. You can set this to `0` if you want to disable this feature altogether, and then sort the footprints by their Wellington scores (e.g. `sort -nk 5 <fp.bed> > <out.bed>`) and choose your threshold this way if you like.

3 DNase-seq footprinting Tutorial

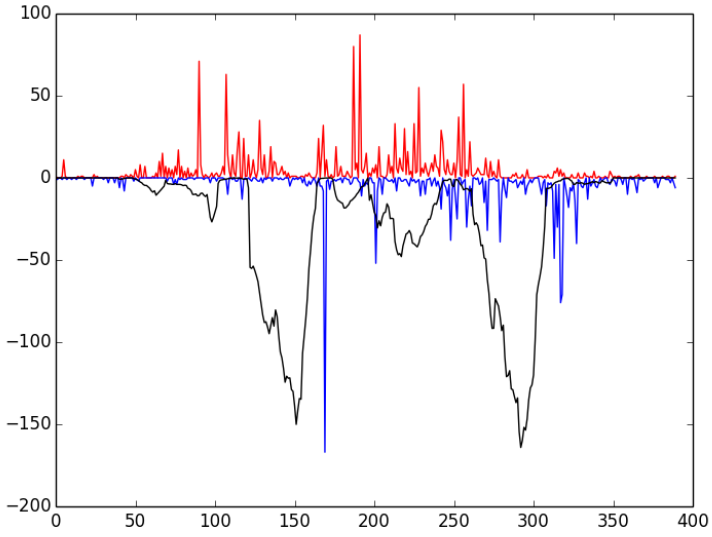
This document gives you a brief outline on how to analyse DNase-seq data. It assumes knowledge of the unerlying biological method, and is meant to help those that generally understand how to use the terminal to perform simple bioinformatic analyses. If you want to brush up on your DNase-seq before continuing, the HOMER docs, the Wellington Paper, the ENCODE DNase-seq paper, are good places to start.

3.1 Testing pyDNase installation

After installing pyDNase, go ahead and run:

```
$ example_footprint_scores.py
```

This script tests that everything has been installed and runs correctly. You should see the following window



If so, congratulations! Everything has installed properly. The red and blue bars correspond to cuts on the positive and negative strand, respectively, and the black line represents the raw Wellington footprint scores.

3.2 Getting data in the correct format

The first, and most important thing you need is the aligned reads from the DNase-seq experiment. If you are working with ENCODE data, you can get the data pre-aligned in `.bam` format from here. Note that you must download the corresponding `.bam.bai` file, which is the index file. You can find Dgf (digital genomic footprinting) reads and hypersensitive sites from the wgEncodeUwDgf folder on the UCSC GoldenPath server. Note that these are different thana those found in the `wgEncodeOpenChromDnase` and `wgEncodeUwDnase` folders, which are not of sufficient sequencing depth to get a good number of footprints from.

Do note that there are two different DNase protocols - the *single-hit* method described by Boyle et al, and the *double-hit* method described by Sabo et al. All of the data generated by ENCODE under the `wgEncodeUwDgf` label all uses the *double-hit* method (the assay is easier to perform and the data is cleaner, so I don't envisage the *single-hit* method coming back). Do not that whilst Wellington can be run on the *single-hit* data, it hasn't been designed to do so.

If you're on the cutting edge, such as generating your own DNase-seq data or using some of the raw (unaligned) data from the NIH roadmap epigenomics project, then you'll need to align the FASTQ files from your sequencer yourself (outlined below).

3.3 Aligning your reads

I used to use Bowtie 1 with the settings (this is basically how all the ENCODE data is aligned):

```
$ bowtie -t -p 8 -v 2 -m 1 --all --best --strata --sam hg19 -f -1 <input.fastq> > <output.sam>
```

But this has several limitations - the suppression of non-uniquely mapping reads angers Lior Pachter, `bowtie` doesn't do well with long reads, and `bowtie` can't handle indels. So where possible, use `bowtie2`, which I usually use with the default settings (example below):

```
$ bowtie2 -x hg19 -t -p 8 -q -U <input.fastq> -S <output.sam>
```

Indels can create 'fake' footprints as they lead to short regions where no sequences can align, so when comparing different samples this becomes important!

3.3.1 Dealing with SRA

If you're getting files from the SRA, you'll need to convert the files from the propietry `.sra` format to `.fastq` `sratoolkit`. Download and install `sratoolkit` from here and, and then use `fastq-dump` to convert to either convert to FASTQ directly, or pipe directly to `bowtie2` such as:

```
$ fastq-dump <reads.sra> -Z | bowtie2 -x hg19 -t -p 8 -q -U - -S <output.sam>
```

3.3.2 Sorting and Indexing

You must then convert these files to sorted, indexed, bam files:

```
$ samtools view -bS <in.sam> > <out.bam>

$ samtools sort <out.bam> <out.sorted>

$ samtools index <out.sorted.bam>
```

At this point you will have <out.sorted.bam> and <out.sorted.bam.bai> - the BAM format is a very common format used for the interchange of aligned sequence data, and lots of common tools like HOMER: and MACS can handle bam files.

Tip: The more technical people will notice than you can pipe directly from `fastq-dump` to `bowtie2` to `samtools view`. The downside being it can be hard to debug problems when you chain a lot of programs together.

3.4 Peak calling

A prerequisite to footprinting the genome is the definition of DNase Hypersensitive Sites (DHSs) - these are regions of the genome where nucleosomes have been displaced and the DNase is free to cut the DNA.

Many peak callers exist such as MACS, MACS2, F-seq, HOMER's FindPeaks, HOTSPOTs (the list is practically endless). There's a good review of peak calling in DNase-seq data here, and identifying DNase hypersensitive sites is outsite of the remit of this tutorial, so I emplore you to read around the area and use your own judgement here.

However, if you *really* want to be spoonfed (gimme the peaks now, I'm in a rush!) then I usually use HOMER's `findPeaks` with the parameters:

```
$ findPeaks -region -size 500 -minDist 50 -o auto -tbp 0
```

converting the HOMER peaks to a BED file using HOMER's builtin `pos2bed.pl` and then merging the overlapping regions with:

```
$ bedtools sort -i <input.bed> | bedtools merge -i - > <output.bed>
```

I find the results are almost exactly the same as the HOTSPOT method employed by ENCODE. See the HOMER: documentation for detailed information on how to carry out this procedure.

3.5 Quick and easy Footprinting

So this is what you're waiting for - *show me the money!* as they say. Armed with your install of pyDNase and your `.bam`, `.bam.bai`, and `.bed` files, you're ready to go! You can go ahead and footprint your DHSs in order to identify protein-DNA binding sites with the following command:

```
$ mkdir K562_footprints
$ wellington_footprints.py K562.DHSs.bed wgEncodeUwDgfK562Aln.bam K562_footprints/
```

By default this will use the number of threads that you have available, on a 16 core machine, this takes about 30 minutes.

You should really take some time to read through the settings in the documentation, you can get this by running:

```
$ wellington_footprints.py -h
```

I often get the comment that footprints from are too stringent. This is a common question - if you have low read depths you might need to adjust the `-fdrlimit` parameter to something less stringent like `"-10"` or `"-5"` (the closer to 0, the more liberal), which sets the mimimum amount of evidence required to support the alternate hypothesis of there being a footprint present.

Tip: You can set `-fdrlimit` to `-0.01` if you want to disable this feature altogether, and then sort the footprints by their Wellington scores (e.g. `sort -nk 5 <fp.bed> > <out.bed>`) and then visualise the footprints choose your threshold this way if you are unsure.

3.6 Interpreting Wellington's Output

Explore the folder that you created above (`K562_footprints`) and you will notice three things.

`wgEncodeUwDgfK562Aln.bam.K562.DHSs.bed.WellingtonFootprints.FDR.0.01.bed` contains the footprints at the FDR of 0.01 - this is a good place to start for your footprints. What is happening here is that the data for each DHS is being randomised, and the p-value cutoff for each DHS is being raised from the baseline of `-fdrlimit` according to how often the random data generate footprints. If you're not happy with the footprints here (i.e. they seem too stringent), then feel free to look at the p-value cutoffs (see below) or rerun with different parameters such as a less stringent `-fdrlimit` (see above).

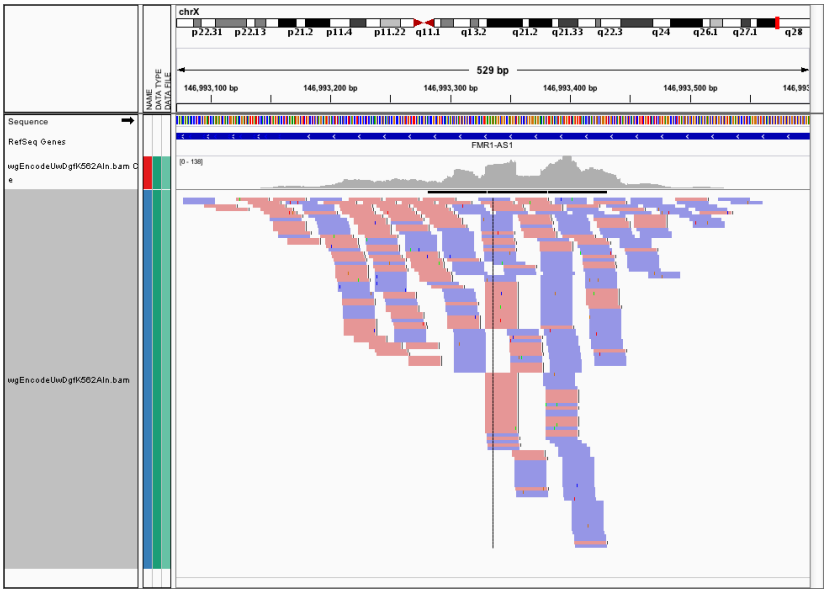
`wgEncodeUwDgfK562Aln.bam.K562.DHSs.bed.WellingtonFootprints.wig` contains the raw footprinting scores - have a look in IGV (you'll need to convert to a bigWig track using UCSC's `wigToBigWig` tool if you've used all the DHSs)

`p value cutoffs` contains the footprints at varying different stringencies - some people prefer this approach to the FDR approach, so these are saved here.

3.7 Visualising the data

You probably want to see what the data looks like. Well you can, with IGV! You can open up the `BED` files (and the `WIG` file) from the output above, and also load up your `.bam` file and have a play around. Have a look at how the different stringencies give you different results.

Often, we want to visualise the raw cut data (just the 5' most ends of the cuts) from a DNase-seq experiment, so visualising the pileups from the `.bam` file isn't helpful here. Here's the FMR1 promoter viewed as a `.bam` file in IGV

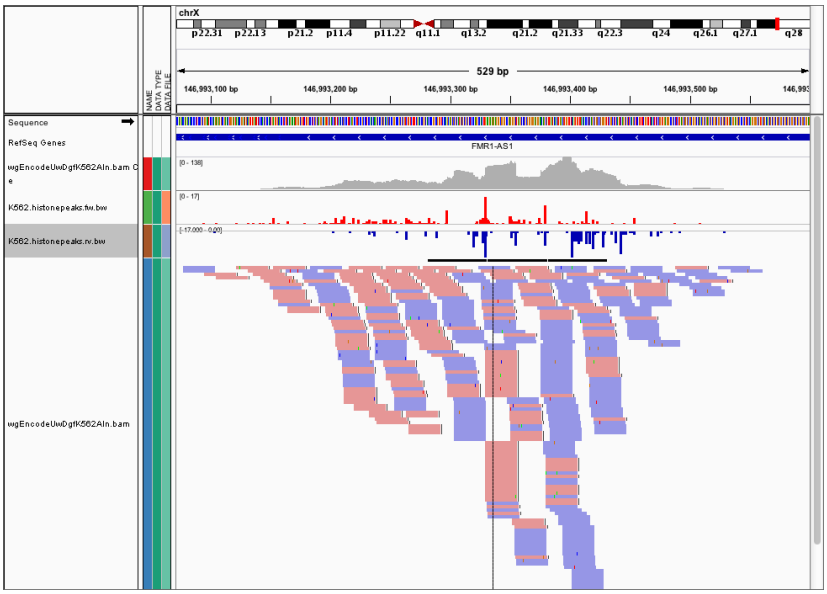


We can use the `dnase_wig_tracks.py` function to generate WIG files based on a BAM file a list of regions of interest. Go ahead and look at the help options for *dnase_wig_tracks.py* and see if you can work out how to generate the wig files and load them in IGV:

```
$ dnase_wig_tracks.py -h
```

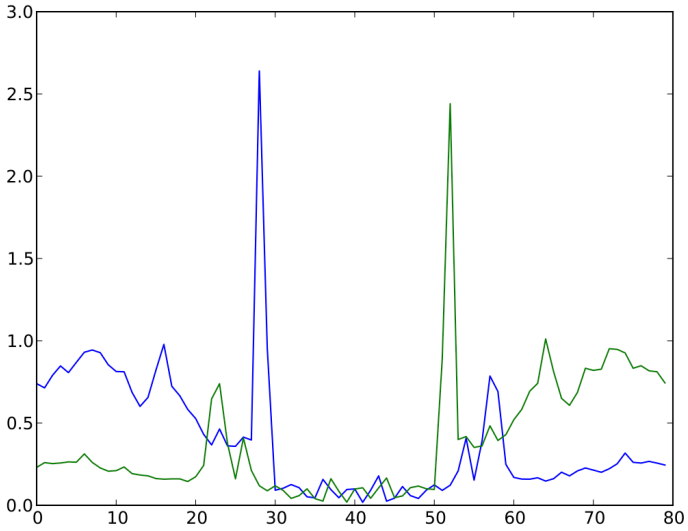
Note: By default, cuts on the reverse strand will be reported as negative numbers (for visualisation). If you want to be using this data for something else, you can pass the `-r` flag, which will use the real number of cuts.

Once you do this, you can load the data into IGV and it should look like this



3.8 Visualising Footprints as average plots

So you have your set of footprints, or your set of footprinted motifs (E-box, CTCF, NFE2, etc...) and you want to see what they look like. Average profile plots illustrating DNase activity surrounding a set of regions are frequently used in papers, like this.



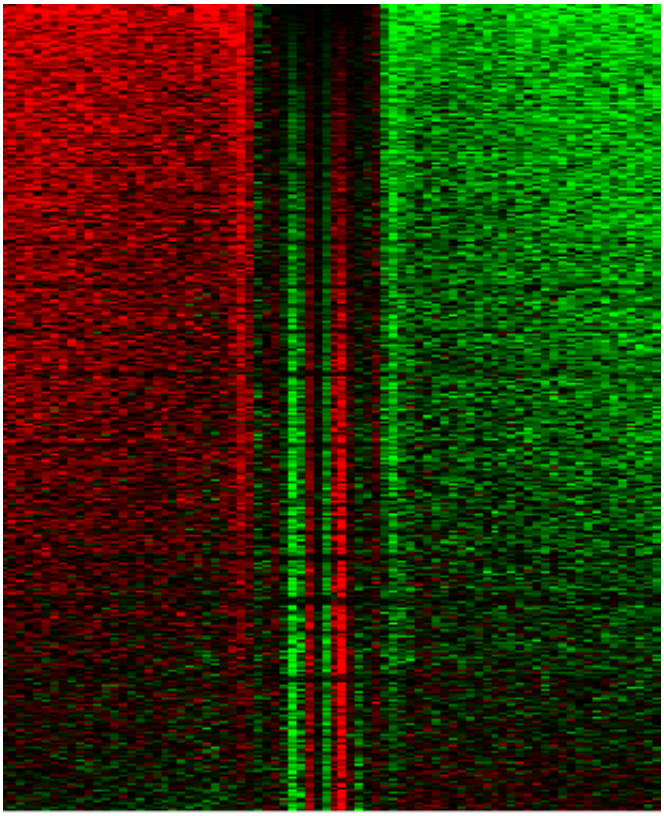
Have a look at the help for `dnase_average_profile.py` and see if you can work out how to display the average profiles for the supplied 3000 K562 CTCF footprinted motifs, `K562_3000_CTCF_Footprints.bed` (or the footprints you discovered earlier). In order to get the locations of specific footprinted transcription factors, you'll need to perform motif finding. Have a play around with the parameters as well:

```
$ dnase_average_profile.py -h
```

Tip: This script uses matplotlib to generate the output, so it will write a filetype based on the file extension provided (e.g. `out.png` or `output.pdf`). Use the file extension you want, and the plot will be generated as that type.

3.9 Visualising Footprints as heat maps

Lots of the time, people don't want averages of the data (like above), but want a heatmap showing the raw data (ideally, combine both in one plot!) like this



Which illustrates footprints for the AP-1 complex in K562 cells. For this you'll need JavaTreeView downloaded. `dnase_to_javatreeview.py` will generate a CSV file that you can put straight into JavaTreeView to visualize your data like above. Once again, go ahead and open up the help:

```
$ dnase_to_javatreeview.py -h
```

You'll notice there are a lot of options here. Go ahead and use some data (e.g. the K562 CTCF footprints or the footprints you calculated earlier) to make a CSV file using the script.

To actually view the data, load up javatreeview using:

```
$ java -Xmx4G -jar TreeView.jar
```

and then use `File->Open`, change the file format box to `All Files` and then chose the CSV file generated from the script above. You'll then want to go into `Settings->Pixel Setting` and check all the `Fill` boxes. Go ahead and play around with the contrast! Play around with the parameters in the `dnase_to_javatreeview.py` script and see how it affects the visualisation.

3.10 Motif Finding

Most of the things that people want to do with their footprints is look for enriched motifs, annotating the nearest TSS to specific factors, etc. I usually use HOMER for this as it fits in with my workflow. We won't have time to go into how to do these analyses here, but HOMER has a really good tutorial on how to find motifs here and has generally very good documentation on annotating genomic regions. I highly recommend you pour yourself a nice glass of wine

and settle down in a fancy leather chair and read the website thoroughly. Don't go running in, guns blazing, running all the tools without understanding all the parameters!

Danger: Make sure when using HOMER's `findMotifsGenome.pl` script, make sure to use the `-size given` parameter or it will just search for all motifs within several hundred basepairs of the footprint!

3.11 Leveraging pyDNase for fun and profit (mainly fun)

If you've survived this far, well done! Fire up the python terminal:

```
$ python
```

And head over to the advanced documentation here and here, which introduces you to how to load up data from a BAM file directly. I don't anticipate many people will get this far, but if you do, I will come and talk you through how to proceed here if you're having trouble following the API specification (although it is a fairly simple API).

Can you answer these questions? If you can't think of how to approach the problem, come and ask me and I'll give you some pointers.

- For the set of 1000 DHSs provided in a bed file - can you work out me the mean number of DNase cuts per DHS?
- Can you plot a histogram of the strand imbalance (the ratio of cuts on the +ve strand to -ve) for these 1000 DHSs?
- Can you write a BED file of these 1000 DHSs annotated with the number of DHS cuts they have in them?

Chapter 4

Discussion

4.1 Footprinting analysis of DNase-seq data

The *Wellington paper* describes the first open-source software implementation of a *de novo* algorithm for predicting transcription factor footprints using DNase-seq data (Wellington). Wellington is designed to detect footprints based on the observed strand imbalance that surrounds known protein-DNA interactions in the double-hit protocol, which increased the predictive performance of DNase-seq footprinting over previous analysis methods [2, 80]. It was hypothesised that the strand imbalance is a natural consequence of the size selection step of the double-hit protocol, which purifies ca. 50-200 base pair DNA fragments produced by DNase I digestion. This was later confirmed by two independent studies that assessed the impact of various experimental parameters in the DNase-seq library preparation protocol [52, 64], highlighting the importance of understanding the biological methods in detail before embarking on the design of analytical techniques.

Wellington was objectively validated against other footprinting and DNase-seq analysis methods using a wide range of techniques, some of which were well-established (i.e. Receiving Operating Characteristic (ROC) analysis using ChIP-seq as a gold standard), along with other metrics such as motif content and phylogenetic conservation. In doing so, concerns were raised with previous validation methods relying almost purely on ChIP-seq recapitulation. A major limitation of validation of transcription factor binding site predictions is the belief that ChIP-seq itself is an objective truth (as discussed in Chapter 1). By designing algorithms to recapitulate ChIP-seq data, the risk of over-fitting data to the small number of ChIP-seq experiments becomes very real. Ny solely evaluating footprinting algorithms on the ability to recapitulate ChIP-seq engenders a blind spot in DNase-seq

footprinting due to reluctancy of predicting a transcription factor binding without a corresponding ChIP-seq result, or vice versa, where there are results from ChIP-seq but there is no corresponding footprint, which could be due to many biological reasons such as transient or indirect binding. Whilst the performance of Wellington has subsequently twice been independently validated ([53, 80]), these authors still solely rely on area AUROC of ChIP-seq recapitulation as a performance measure, which heavily favours the correct prediction of True Negatives over False Positives in DNase-seq footprinting studies. Even though the AUROC can be misleading with regards to the performance of a predictor, analysis of the shape of the ROC curve can be informative to the performance characteristic over the entire sensitivity range, i.e. the initial slope of the ROC curve is indicative of the PPV of a footprinting algorithm.

It was suggested in this publication that ENCODE's claim that there are 0.4 to 2.3 million genomic footprints, dependent on the cell type [66], was exaggerated, due to the large number of motif-less and low conservation scoring false positives present in the ENCODE predictions. Others have also commented on the exaggeration of ENCODE's claims of 'functionality' in the human genome due to the distinct lack of sequence conservation in 'functional' elements [69], a finding which recapitulates the discovery presented here e.g. a large set of the ENCODE footprints are not conserved and have low motif content. However, it remains difficult to determine exactly how many binding sites there may actually be in a given DNase-seq dataset, and therefore the human genome, as DNase-seq footprinting in humans is still limited by sequencing depth[2, 66]. As the cost of sequencing continues to decrease, the predictive power of DNase-seq footprinting techniques may eventually reach a limit.

Significantly reducing the number of false-positive predictions and increasing the positive predictive power gained from DNase-seq has the largest impact on assays that have no technical or biological replicates, which is often the case with primary tissues such as patient samples – even the ENCODE DNase-seq data performed on cell lines lack biological or technical replicates. Ultimately, the preference for high specificity over high sensitivity depends on the analysis being performed, but by allowing the user of the tool to alter the stringency of the parameters, this decision can be based on the performance characteristics of the results that they find acceptable for their downstream analyses.

The setting of default parameters for Wellington was challenging. Whilst scripts to perform the analysis along with the parameters used in the paper were provided, the reviewers insisted that the method should be a 'point and click'

exercise, without the need for the user to specify default parameters. Such a philosophy all too easily leads to overfitting, as all the data analysed here was generated in one laboratory — and there was no evidence that the same parameters would be suitable for data generated by others. Because of the setting of default parameters, a large number of users in other labs have commented on the stringency of the default parameters. This issue is not too dissimilar to the problems that arose with the popular aligner, Bowtie [81]. In Bowtie 1, there were no default parameters and the user had to read extensive documentation in order to perform an alignment, which led to user confusion. In Bowtie 2 [82], the developers provide a default set of parameters so the tool can be run easily, but without the user understanding the parameters behind the software. As such, many of the emails from people would be easily answered if the users had read the paper and documentation and understood the role of each parameter in the Wellington algorithm.

In the remainder of the thesis, the underlying Wellington method remains unchanged apart from implementation changes to increase computational speed. It remains to be seen how footprinting algorithms can be further enhanced. Even though it is known that the pattern of the DNase-seq signal surrounding protein-DNA binding events is transcription factor-dependent [53, 66, 67], Wellington outperforms CENTIPEDE, a method designed to use the transcription factor-specific footprint patterns, even though a single model to search for all possible transcription factor-binding events in a DNase-seq data set is used. Two further footprinting methods have been published since the publication of the Wellington paper: PIQ [83] and DNase2TF [53]. PIQ appears to have more false positives than true positives, but because of the Area under the ROC (AUROC) statistic used to validate this, the number of true negatives can create the impression that this method performs adequately. DNase2TF, whilst using a much more complicated model that incorporates corrections for the DNase I cutting bias (the data for which were not available when Wellington was developed) only yields moderate improvements over Wellington using the AUROC statistic (discussed further in Section 4.4).

To encourage further investigations, pyDNase and Wellington were released as a Python package for the fast and easy analysis of DNase-seq data. It was hoped that this would accelerate both the analysis of DNase-seq data and the development of further footprinting algorithms. It has been reassuring to observe that in just over 1 year since publication, other research groups are benchmarking new analyses against Wellington [53, 80], and applying the Wellington to biological

data in a research article [84]. Additionally, several researchers have even taken time to report bugs and provide code fixes to the pyDNase github page.

4.2 Application of Wellington to clinical samples

Shortly after the development of Wellington, it was utilised in an integrative genomics project aimed (Appendix A) at identifying the core transcriptional network regulating self-renewal and differentiation block in t(8;21) acute myeloid leukaemia (AML) [85]. The t(8;21) translocation is a common chromosomal rearrangement that account for 10% of diagnosed AML [86, 87], giving rise to the RUNX1/ETO fusion protein. Here the DNA binding domain of RUNX1, a transcriptional activator essential for haematopoietic identity [88] becomes fused to the transcriptional repressor ETO. This fusion leads to the formation of an aberrant transcription factor. The RUNX1/ETO fusion protein binds RUNX1 DNA motifs (TGYGGT) in the genome, but leads to systemic transcriptional repression rather than transcriptional activation [89].

Wellington was used to investigate the similarity of RUNX1/ETO binding characteristics between the t(8;21) Kasumi-1 cell line and two patient-derived samples in order to validate Kasumi-1 cells as a valid model system. This was illustrated through the use of footprinting analysis of peaks found in the intersection of RUNX1/ETO, RUNX1, LMO2, and HEB ChIP-seq experiments from the Kasumi-1 cell line. Motif searching was then performed for the ETS, RUNX1, and HEB (E-box) motifs, where it was that Wellington was accurately predicting these transcription factor binding sites in the clinical samples, and that all four motif show footprints, illustrating that all proteins are directly in contact with the DNA here, and that the clinical sample show similar binding characteristics to the cell line. Footprinting DNase-seq data from the clinical sample within RUNX1/ETO peaks from Kasumi-1 cells revealed co-localisation of PU.1, RUNX1, ERG, and SCL motifs in these peaks via a co-occurrence analysis. Here, using one DNase-seq experiment, we correctly identified RUNX1/ETO binding partners.

This study demonstrated the ability of DNase-seq footprinting analysis as replacement for ChIP-seq for the identification of transcription factor binding sites by recapitulating the known transcriptional complex of PU.1, RUNX1, ERG, and SCL through the complementary use of RUNX1/ETO ChIP-seq with DNase-seq footprinting to identify binding partners. The validation of this method on clinical samples is of great importance, where it is not feasible to perform multiple ChIP-seq experiments due the large volumes of cellular material needed for sev-

eral immunoprecipitations; this is the first study to report successful DNase-seq footprinting on clinical samples.

From a computational standpoint, it was reassuring that the Wellington method performed adequately on data generated externally of ENCODE. As the methods had differing starting material of clinical samples (which are much more heterogenous than the cell lines used in ENCODE), and a different DNase digestion protocol, with the removal of the nuclear isolation stage used in the ENCODE protocol. Nevertheless, the default Wellington parameters used in Chapter 2 were adequate. Unfortunately, differential DNase-seq footprinting, as described in Chapter 3, was not available for this paper, as this would have helped link the binding of specific transcription factors to changes in gene expression. Overall, this study demonstrated of the power of using footprinting in a functional genomics focussed analysis of a biological system, augmenting several other assays in order to provide novel biological insights.

4.3 Differential DNase-seq footprinting

The *Differential Footprinting* paper introduces Wellington-bootstrap, an extension to Wellington that allows for the identification of differential footprints between two datasets. This addresses the unmet need for a method that can identify differences in DNase-seq footprints between two datasets beyond performing simple present/absent footprinting comparisons. The method itself follows a conceptually simple approach: Wellington was used to identify a footprint in a dataset, and it was then determined whether the second dataset recapitulates this footprint by comparing combined data from a second dataset to that where data from the second dataset was randomly shuffled.

This assumes that the Wellington parameters (i.e. the footprint and shoulder regions) between the two datasets are comparable. Whilst more complicated analyses that accounted for the ability for footprints to either grow or shrink in size, or even shift several basepairs, were considered, it was discovered that this did not yield a difference in the results, it led to a ca. 10× increase in the time taken to perform an analysis due to the additional parameters in the resulting model. It could be that events such as footprint size changes and footprint shifting are rare, undetectable by DNase-seq, or do not occur in a large number within the datasets that were analysed here.

Whilst the DNase I cutting bias has been thoroughly quantified since the publication of the *Wellington* paper [51–54] it was decided not to alter the Wellington

method to account for the DNase I cutting bias when calculating Wellington scores. Whilst there have since been two novel footprinting methods (DNase2TF [53], and an unnamed method with no software implementation [54]) that accounts for the DNase I cutting bias, the authors did not perform a comprehensive analysis of how much the bias correction contributes to their footprinting performance independent of their choice of statistical model. Because of this, a comprehensive analysis of how bias correction affects footprinting performance is warranted before any alteration of the underlying Wellington method.

A limitation of the Wellington-bootstrap method is that the method only performs comparisons between two datasets, and the method had to be applied in a pairwise fashion over 7 datasets in order to gain an understanding of the transcription factor specificity. However, as this approach has time complexity of $\mathcal{O}(n^2)$, at roughly 192 CPU hours per comparison on a 2.66 Ghz Intel Xeon (Nehalem) processor, the computational time required becomes prohibitive with an increase in the number of datasets being analysed. Computational as well as analytical challenges must be solved simultaneously in order to gain the most power out of comparative DNase-seq footprinting studies. Considering the competitiveness of the development of DNase-seq footprinting strategies (with at least 8 analysis methods described in the last 5 years), there has been little focus on the speed of the implementation compared to the scientific validation apart from speed benchmarks released in the latest method, DNase2TF [53]. As part of the *Differential Footprinting* paper, significant updates to the underlying pyDNase library were released (pyDNase 0.2.0). Here, larger portions of the Wellington algorithm were written in C (140 lines of code in C in the original Wellington vs 399 lines of code in C in pyDNase 0.2.0). Additionally, the code was parallelised using python's multiprocessing module — on a desktop machine with two 2.66 Ghz quad core Intel Xeon (Nehalem) processors. This reduced the tine required to calculate the Wellington scores on a single dataset from around 24 hours to about 30 minutes.

Footprinting multiple DHSs is an 'embarrassingly parallel'[1] problem, and therefore scales virtually linearly with the number of CPUs dedicated to the task. A 'proof of concept' version of Wellington that used OpenMPI for use on a cluster environment allowed the computation of results in 5 minutes by leveraging hundreds of cores at once. The OpenMPI implementation was not published alongside this paper, though, as high-performance compute environments are usually highly bespoke, and it would have been problematic to support an OpenMPI version. It is envisaged that a parallelised iPython version will be released to make use of

[1]The job can easily be split into one task per DHS.

cluster computer environments, as the underlying iPython parallelisation engine provides a backend to interface with many popular job scheduling environments.

The Wellinton-bootstrap method opens the door to be able to build better models for gene regulation by comparing changes in transcription factor occupancy on a massively parallel level that is only possible with DNase-seq footprinting. Attempting to use ChIP-seq based studies to perform a similar analysis would not only be economically unfeasible at the current cost of high-throughput sequencing, but would require knowledge of all the transcription factors of interest.

The results gained from the comparisons of the data from 7 clinical samples from the NIH Roadmap epigenomics project [90] were only possible by DNase-seq footprinting with Wellington. Our paper describes the first DNase-seq footprinting analysis on cells from healthy donor, and in addition to 5 haematopoietic cell types that were isolated via Fluorescence Activated Cell Sorting (FACS) from the same healthy donor. Using Wellington Bootstrap, it was possible to differentiate between the isolated populations purely based on transcription factor motif content within differential footprints, with minimal prior knowledge of the system.

It was also shown that changes in footprinting status in gene promoters can be linked to changes in gene expression at neighbouring genes, where using differential DHS scores linked with a motif is unable to do so. This is most likely due to the large size of the DNase hypersensitive regions (up to 2000bp) and the transcription factor motif-rich landscape of gene promoters providing many false positives when performing motif searches in promoters without the use of footprinting analyses. Combined with chromatin conformation capture techniques, this method could be used to link the binding of transcription factors at enhancers to gene expression without the need for antibodies. Alternatively, another approach that would be interesting would be a similar analysis to the DNaseI sensitivity quantitative trait locus (dsQTL) approach [33], but instead of linking DHS strength to gene expression using a population study, linking footprint occupancy to gene expression across the genome. A weakness of this study was the lack of experimental validation of the predicted differential binding sites. ChIP-seq data for a set of transcription factors across several of the clinical samples may allow the rigorous validation of the predictions, providing better estimates of the sensitivity and specificity of the method.

4.4 Outlook

At the onset of this project, DNase-seq footprinting was a technique with only a handful of documented analysis approaches, almost none of which had software implementations, and those that did required considerable of data processing from the raw sequencing data into non-standardised file formats in order to perform analyses. Wellington was introduced as a leader in the DNase-seq footprinting field. Not only did it objectively perform better over a wide range of performance metrics, which has subsequently been independently validated [80]. Wellington was distributed as an easy-to-use software package that did not rely on non-standardised file formats, instead opting to use the industry standard BAM file format for sequence alignments [91]. Wellington was built on top of the pyDNase software library, a Python library that allows for the easy interaction with DNase-seq data, containing several convenience scripts that demonstrate the power of the pyDNase library along with a DNase-seq footprinting tutorial.

The development of novel DNase-seq footprinting analysis methods has been steadily increasing, with several unpublished (MOCCA, DNaseR) and published (PIQ [83], DNase2TF [53]) methods that perform DNase-seq footprinting having been released since the publication of the *Wellington* paper. However, the development of new footprinting methods has proven to be distinctly less competitive than other functional genomic studies such as ChIP-seq and RNA-seq. This is probably due to the relatively small number of groups able to perform the DNase-seq experimental protocol, which is inherently more difficult to perform (and analyse) than ChIP-seq [52, 92]. In addition to the technical challenges involved in performing DNase-seq experiments, the cost of the method is also prohibitive, with samples with read counts of up to 1.2 billion reads per sample still yielding increasing numbers of footprints — it is hard to estimate at what read depth the number of footprints detected will plateau. The risk inherent in performing DNase-seq is therefore seen as relatively high, especially given that the analysis methods are relatively immature compared to ChIP-seq.

There are two distinct approaches to footprinting present in the literature, those that partition all occurrences of a motif in a genome into bound and unbound states, and those that operate directly on the data with no prior knowledge. Both of these approaches seek to answer different questions, although the partitioning approach is easier to validate due to being easily comparable to ChIP-seq. However, not enough distinction is made in the literature between true *de novo* footprinting algorithms (The Hesselberth method [61], the Neph

Method [66], Wellington [2], DNaseR (unpublished: `http://www.bioconductor.org/packages/release/bioc/html/DNaseR.html`) and DNase2TF [53]) and motif-centric partitioning approaches (CENTIPEDE [67], MOCCA (unpublished: `https://github.com/ajank/mocca`), PIQ [83]).

Additionally, some of these footprinting methods were designed for and only tested on the original single-hit protocol (CENTIPEDE, Hesselberth method), whilst the others were designed around and only tested on the double-hit protocol (Neph method, DNase2TF). It appears that only Wellington and MOCCA have been comprehensively tested against both methods at the time of publication. A recent study highlighted the effect of library preparation protocols on the quality of sequencing data [52], illustrating the importance that analytical techniques need to be developed alongside new experimental methods in order to gain the greatest predictive accuracy power from DNase-seq footprinting. A potential avenue of interest for further algorithmic development is the assessment of allelic differences in TF binding. Even though the ability to detect allelic transcription factor footprints has been demonstrated by utilising the underlying sequence data from DNase-seq to detect heterozygous SNPs and subsequently phase the the alignments [66], no footprinting models currently take allelic differences into account when performing footprinting. Where variants exist that disrupt transcription factor binding, analysing the data on an allele-by-allele basis would prevent a false negative prediction in the cases where combining the data from the alleles conceals a footprint, and could provide an estimate to the number of allelic transcription factor binding sites within the genome.

From the benchmarking performed in the *Wellington* paper, it appears that for some transcription factors the performance of almost all analytic approaches are reaching a common upper limit of predictive accuracy. This limit is also seen in other benchmarking efforts [53, 80]. This saturation phenomenon could highlight either the unsuitability of ChIP-seq data to be a 'gold standard', or that all analyses are reaching a upper bound on the predictive power of DNase-seq data. It remains to be seen if further refinements to the experimental protocol, such as the use of novel nucleases such as benzonase and cyanase [64] will 'unlock' further predictive power of DNase-seq footprinting.

Another possible explanation for the discrepancies between DNase-seq footprinting and ChIP-seq data is the reduced ability for DNase-seq to recapitulate more transient protein-DNA interactions due to the lack of the cross linking present in the ChIP-seq protocol, which can capture much shorter interactions. It has been suggested that shorter binding time of a transcription factor to DNA leads

to 'shallower' footprints, [53], but this study was limited to only select transcription factors, and the cutting bias on DNase I was not taken into account when making this observation, so it remains to be seen if this is true for all transcription factors. Further time course assays would help shed light in this area, revealing the time resolution of DNase-seq footprinting by capturing the dynamics of the transcription factor binding site landscape over time.

With DNase-seq being a recent and relatively expensive assay, the lack of any replicates in the ENCODE Digital Genomic Footprinting DNase-seq data makes quantifying the reliability of the method difficult, has not allowed the quantification of technical and biological variation in the assays. However, the DNase-seq data released as part of the NIH Roadmap Epigenomics Consortium [93] exists for specific subpopulations of cells for a number of individuals. Analysis of the data could assess the consistency of the chromatin and transcription factor landscape for the same cell population between individuals (i.e. biological replicates).

A major limitation preventing the widespread adoption of DNase-seq is the inability of the method to scale to a low (or even single) cell level, as with ChIP-seq [94]. This is because a maximum of two fragments (for diploid cells) can be generated per DHS per cell. It therefore becomes impossible to resolve the tissue heterogeneity problem in DNase-seq footprinting — does a 'weak' footprint in DNase-seq data correspond to weak binding in the entire population of cells in the assay, or to strong binding in a small subset? However, an improvement in this area might be on the horizon through the use of another recently described method used to map chromatin accessibility. The assay for transposase-accessible chromatin by sequencing (ATAC-seq) that uses the ability of a tranposase to integrate into DNA as a measure for DNA accessibility and can be performed on as little as 500 cells [95]. Currently, it is uncertain whether it is possible to analyse the resulting data to identify protein-DNA interactions at the same scale that can be achieved with DNase-seq data.

Beyond the development of new analytical models for footprinting DNase-seq data, other computational challenges, specifically software engineering hurdles, have the possibility of being a major limiting factor to further innovations in the area. To allow for more comprehensive comparative analyses between large numbers of experiments to take place (i.e. to perform a similar analysis as undertaken in the *Differential Footprinting* paper for all 41 ENCODE samples), computational efficiency of the underlying methods becomes important. A recent integrative analysis of DNase-seq experiments in assessing the impact of DHSs on gene expression [33] had to implement an entire custom pipeline for their analysis (and neglected to

publish the code behind it), which illustrates the barrier of entry for these methods to become mainstream. DNase2TF claims to be the fastest method for footprinting, with a single dataset only taking around 30 minutes to analyse. The usefulness of their method is somewhat detracted by their use of non open-source software (MATLAB), the lack of software documentation, and the necessity to process files into non-standardised file format. In the updates introduced to Wellington as part of the *Differential Footprinting* paper, Wellington's run time is now of comparable magnitude.

4.5 Conclusions

The results presented in this thesis represent significant advances in the prediction of transcription factor binding sites from DNase-seq data, and provide free and open source software tools and comprehensive documentation that facilitate the application of the Wellington and Wellington-bootstrap methods. It is hoped that through the increased specificity of DNase-seq footprinting alongside demonstrations of the power of the method in clinical research scenarios inspire continued research into the development of DNase-seq footprinting methods and stimulates interest in the biological insights the data have to offer.

Because of the publication of the Wellington algorithm as an easy-to-use and well-documented software tool, the ability for others to incorporate DNase-seq footprinting analyses into their research is no longer limited to those that have extensive expertise in DNase-seq signal processing. DNase-seq footprinting complements other genomic assays in a number of ways: by providing information about the transcriptional regulation landscape without having to specify interest in a factor *a priori* [66, 84], by supplementing ChIP-seq experiments in order to differentiate direct from indirect binding[85], or as a direct surrogate for ChIP-seq by integrating PWM data [67, 83].

Transcription factors remain attractive targets for drug design due their involvement with almost all cellular processes [96], and understanding their behaviour is therefore pivotal in the ability to construct models that explain gene expression. Even though transcriptional regulation becomes aberrant in a large number of cancers [97, 98], only three transcription factor families are currently targeted by cancer treatments — thiazolidinediones target the retinoid X receptors (RXRs), tamoxifen targets the oestrogen receptor, and alitretinoin targets the retinoic acid receptors (RARs) [99]. In the *Differential Footprinting* paper, differential footprinting is used in order to reveal the transcription factors driving

cell identities. However, interpreting differential occupancy patterns will not only allow for the identification of the transcription factor networks driving healthy cell types, but also distinguish the events that differentiate normal from diseased cellular states. Moreover, these analyses will provide evidence for certain factors' involvement in specific cancers, which in turn, will identify disease-specific transcription factors and signalling pathways as targets for therapeutic interventions in stratified approaches to treating cancer.

Bibliography

[1] Alan P Boyle, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, January 2008.

[2] J Piper, M C Elze, P Cauchy, P N Cockerill, C Bonifer, and S Ott. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic acids research*, September 2013.

[3] P J Mitchell and R Tjian. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science (New York, NY)*, 245(4916): 371–378, July 1989.

[4] J D Parvin and P A Sharp. DNA topology and a minimal set of basal factors for transcription by RNA polymerase II. *Cell*, 73(3):533–540, May 1993.

[5] M Ptashne and A Gann. Transcriptional activation by recruitment. *Nature*, 386(6625):569–577, April 1997.

[6] F Rojo. Mechanisms of transcriptional repression. *Current opinion in microbiology*, 4(2):145–151, April 2001.

[7] CD Allis and SM Gasser. Chromosomes and expression mechanisms New excitement over an old word: 'chromatin'. *Current opinion in genetics & development*, 8(2):137–139, April 1998.

[8] The ENCODE Project Consortium, The ENCODE Project Consortium, Overall coordination data analysis coordination, Data production leads data production, Lead analysts data analysis, Writing group, NHGRI project management scientific management, Principal investigators steering committee, Boise State University and University of North Carolina at Chapel Hill Proteomics groups (data production and analysis), Broad Institute Group (data

production and analysis), Cold Spring Harbor, University of Geneva, Center for Genomic Regulation, Barcelona, RIKEN, Sanger Institute, University of Lausanne, Genome Institute of Singapore group (data production and analysis), Data coordination center at UC Santa Cruz (production data coordination), Duke University, EBI, University of Texas, Austin, University of North Carolina-Chapel Hill group (data production and analysis), Genome Institute of Singapore group (data production and analysis), HudsonAlpha Institute, Caltech, UC Irvine, Stanford group (data production and analysis), Lawrence Berkeley National Laboratory group targeted experimental validation, data production, NHGRI groups analysis, Sanger Institute, Washington University, Yale University, Center for Genomic Regulation, Barcelona, UCSC, MIT, University of Lausanne, CNIO group (data production and analysis), Stanford-Yale, Harvard, University of Massachusetts Medical School, University of Southern California/UC Davis group (data production and analysis), University of Albany SUNY group (data production and analysis), University of Chicago, Stanford group (data production and analysis), University of Heidelberg group (targeted experimental validation), University of Massachusetts Medical School Bioinformatics group (data production and analysis), University of Massachusetts Medical School Genome Folding group (data production and analysis), University of Washington, University of Massachusetts Medical Center group (data production and analysis), and Data Analysis Center (data analysis). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 488(7414):57–74, April 2013.

[9] Nicole Happel and Detlef Doenecke. Histone H1 and its isoforms: contribution to chromatin structure and function. *Gene*, 431(1-2):1–12, February 2009.

[10] Wikimedia Commons. File:chromatin structures.png — wikimedia commons, the free media repository, 2014. URL `\url{http://commons.wikimedia.org/wiki/File:Chromatin_Structures.png}`. [`http://commons.wikimedia.org/wiki/File:Chromatin_Structures.png`; accessed 17-December-2014].

[11] S Ogbourne and T M Antalis. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *The Biochemical journal*, 331 (Pt 1):1–14, April 1998.

[12] Kenneth S Zaret and Jason S Carroll. Pioneer transcription factors: establish-

ing competence for gene expression. *Genes & development*, 25(21):2227–2241, November 2011.

[13] Peter N Cockerill. Structure and function of active chromatin and DNase I hypersensitive sites. *The FEBS journal*, 278(13):2182–2210, July 2011.

[14] M Mohrs, C M Blankespoor, Z E Wang, G G Loots, V Afzal, H Hadeiba, K Shinkai, E M Rubin, and R M Locksley. Deletion of a coordinate regulator of type 2 cytokine expression in mice. *Nature immunology*, 2(9):842–847, September 2001.

[15] François Spitz, Federico Gonzalez, and Denis Duboule. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell*, 113(3):405–417, May 2003.

[16] Len A Pennacchio, Wendy Bickmore, Ann Dean, Marcelo A Nobrega, and Gill Bejerano. Enhancers: five essential questions. *Nature reviews. Genetics*, 14(4):288–295, April 2013.

[17] Nathaniel D Heintzman, Gary C Hon, R David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F Harp, Zhen Ye, Leonard K Lee, Rhona K Stuart, Christina W Ching, Keith A Ching, Jessica E Antosiewicz-Bourget, Hui Liu, Xinmin Zhang, Roland D Green, Victor V Lobanenkov, Ron Stewart, James A Thomson, Gregory E Crawford, Manolis Kellis, and Bing Ren. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, May 2009.

[18] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K Canfield, Morgan Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Erika Giste, Audra K Johnson, Ericka M Johnson, Tanya Kutyavin, Bryan Lajoie, Bum-Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Alexias Safi, Minerva E Sanchez, Amartya Sanyal, Anthony Shafer, Jeremy M Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari, Michael O Dorschner, R Scott Hansen, Patrick A Navas, George Stamatoyannopoulos, Vishwanath R Iyer, Jason D Lieb, Shamil R Sunyaev, Joshua M Akey, Peter J Sabo, Rajinder Kaul, Terrence S Furey,

Job Dekker, Gregory E Crawford, and John A Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, September 2012.

[19] Amartya Sanyal, Bryan R Lajoie, Gaurav Jain, and Job Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, September 2012.

[20] Y Huang, S J Myers, and R Dingledine. Transcriptional repression by REST: recruitment of Sin3A and histone deacetylase to neuronal genes. *Nature neuroscience*, 2(10):867–872, October 1999.

[21] Suresh Cuddapah, Raja Jothi, Dustin E Schones, Tae-Young Roh, Kairong Cui, and Keji Zhao. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Research*, 19(1):24–32, January 2009.

[22] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, October 2004.

[23] ENCODE Project Consortium, Ewan Birney, John A Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R Gingeras, Elliott H Margulies, Zhiping Weng, Michael Snyder, Emmanouil T Dermitzakis, Robert E Thurman, Michael S Kuehn, Christopher M Taylor, Shane Neph, Christoph M Koch, Saurabh Asthana, Ankit Malhotra, Ivan Adzhubei, Jason A Greenbaum, Robert M Andrews, Paul Flicek, Patrick J Boyle, Hua Cao, Nigel P Carter, Gayle K Clelland, Sean Davis, Nathan Day, Pawandeep Dhami, Shane C Dillon, Michael O Dorschner, Heike Fiegler, Paul G Giresi, Jeff Goldy, Michael Hawrylycz, Andrew Haydock, Richard Humbert, Keith D James, Brett E Johnson, Ericka M Johnson, Tristan T Frum, Elizabeth R Rosenzweig, Neerja Karnani, Kirsten Lee, Gregory C Lefebvre, Patrick A Navas, Fidencio Neri, Stephen C J Parker, Peter J Sabo, Richard Sandstrom, Anthony Shafer, David Vetrie, Molly Weaver, Sarah Wilcox, Man Yu, Francis S Collins, Job Dekker, Jason D Lieb, Thomas D Tullius, Gregory E Crawford, Shamil Sunyaev, William S Noble, Ian Dunham, France Denoeud, Alexandre Reymond, Philipp Kapranov, Joel Rozowsky, Deyou Zheng, Robert Castelo, Adam Frankish, Jennifer Harrow, Srinka Ghosh, Albin Sandelin, Ivo L Hofacker, Robert Baertsch, Damian Keefe, Sujit Dike, Jill

Cheng, Heather A Hirsch, Edward A Sekinger, Julien Lagarde, Josep F Abril, Atif Shahab, Christoph Flamm, Claudia Fried, Jörg Hackermüller, Jana Hertel, Manja Lindemeyer, Kristin Missal, Andrea Tanzer, Stefan Washietl, Jan Korbel, Olof Emanuelsson, Jakob S Pedersen, Nancy Holroyd, Ruth Taylor, David Swarbreck, Nicholas Matthews, Mark C Dickson, Daryl J Thomas, Matthew T Weirauch, James Gilbert, Jorg Drenkow, Ian Bell, XiaoDong Zhao, K G Srinivasan, Wing-Kin Sung, Hong Sain Ooi, Kuo Ping Chiu, Sylvain Foissac, Tyler Alioto, Michael Brent, Lior Pachter, Michael L Tress, Alfonso Valencia, Siew Woh Choo, Chiou Yu Choo, Catherine Ucla, Caroline Manzano, Carine Wyss, Evelyn Cheung, Taane G Clark, James B Brown, Madhavan Ganesh, Sandeep Patel, Hari Tammana, Jacqueline Chrast, Charlotte N Henrichsen, Chikatoshi Kai, Jun Kawai, Ugrappa Nagalakshmi, Jiaqian Wu, Zheng Lian, Jin Lian, Peter Newburger, Xueqing Zhang, Peter Bickel, John S Mattick, Piero Carninci, Yoshihide Hayashizaki, Sherman Weissman, Tim Hubbard, Richard M Myers, Jane Rogers, Peter F Stadler, Todd M Lowe, Chia-Lin Wei, Yijun Ruan, Kevin Struhl, Mark Gerstein, Stylianos E Antonarakis, Yutao Fu, Eric D Green, Ulaş Karaöz, Adam Siepel, James Taylor, Laura A Liefer, Kris A Wetterstrand, Peter J Good, Elise A Feingold, Mark S Guyer, Gregory M Cooper, George Asimenos, Colin N Dewey, Minmei Hou, Sergey Nikolaev, Juan I Montoya-Burgos, Ari Löytynoja, Simon Whelan, Fabio Pardi, Tim Massingham, Haiyan Huang, Nancy R Zhang, Ian Holmes, James C Mullikin, Abel Ureta-Vidal, Benedict Paten, Michael Seringhaus, Deanna Church, Kate Rosenbloom, W James Kent, Eric A Stone, NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Serafim Batzoglou, Nick Goldman, Ross C Hardison, David Haussler, Webb Miller, Arend Sidow, Nathan D Trinklein, Zhengdong D Zhang, Leah Barrera, Rhona Stuart, David C King, Adam Ameur, Stefan Enroth, Mark C Bieda, Jonghwan Kim, Akshay A Bhinge, Nan Jiang, Jun Liu, Fei Yao, Vinsensius B Vega, Charlie W H Lee, Patrick Ng, Atif Shahab, Annie Yang, Zarmik Moqtaderi, Zhou Zhu, Xiaoqin Xu, Sharon Squazzo, Matthew J Oberley, David Inman, Michael A Singer, Todd A Richmond, Kyle J Munn, Alvaro Rada-Iglesias, Ola Wallerman, Jan Komorowski, Joanna C Fowler, Phillippe Couttet, Alexander W Bruce, Oliver M Dovey, Peter D Ellis, Cordelia F Langford, David A Nix, Ghia Euskirchen, Stephen Hartman, and Al... Urban. Identification and analysis of functional elements

in 1human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, June 2007.

[24] Toshiyuki Shiraki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, Akira Watahiki, Mari Nakamura, Takahiro Arakawa, Shiro Fukuda, Daisuke Sasaki, Anna Podhajska, Matthias Harbers, Jun Kawai, Piero Carninci, and Yoshihide Hayashizaki. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15776–15781, December 2003.

[25] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, July 2008.

[26] Zhenhai Zhang and B Franklin Pugh. High-resolution genome-wide mapping of the primary structure of chromatin. *Cell*, 144(2):175–186, January 2011.

[27] Paul G Giresi, Jonghwan Kim, Ryan M McDaniell, Vishwanath R Iyer, and Jason D Lieb. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, 17(6):877–885, June 2007.

[28] Juan M Vaquerizas, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, 10(4):252–263, April 2009.

[29] Alexander Gusev, S Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, Schizophrenia Working Group of the Psychiatric Genomics Consortium, SWE-SCZ Consortium, Anna K Kähler, Christina M Hultman, Shaun M Purcell, Steven A McCarroll, Mark Daly, Bogdan Pasaniuc, Patrick F Sullivan, Benjamin M Neale, Naomi R Wray, Soumya Raychaudhuri, Alkes L Price, Schizophrenia Working Group of the Psychiatric Genomics Consortium, and SWE-SCZ Consortium. Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *American journal of human genetics*, 95(5):535–552, November 2014.

[30] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K Johnson, Theresa K Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R Scott Hansen, Shane Neph, Peter J Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R Sunyaev, Rajinder Kaul, and John A Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science (New York, NY)*, 337(6099):1190–1195, September 2012.

[31] Shyam Prabhakar, James P Noonan, Svante Pääbo, and Edward M Rubin. Accelerated evolution of conserved noncoding sequences in humans. *Science (New York, NY)*, 314(5800):786, November 2006.

[32] Felicity C Jones, Manfred G Grabherr, Yingguang Frank Chan, Pamela Russell, Evan Mauceli, Jeremy Johnson, Ross Swofford, Mono Pirun, Michael C Zody, Simon White, Ewan Birney, Stephen Searle, Jeremy Schmutz, Jane Grimwood, Mark C Dickson, Richard M Myers, Craig T Miller, Brian R Summers, Anne K Knecht, Shannon D Brady, Haili Zhang, Alex A Pollen, Timothy Howes, Chris Amemiya, Jen Baldwin, Toby Bloom, David B Jaffe, Robert Nicol, Jane Wilkinson, Eric S Lander, Federica Di Palma, Kerstin Lindblad-Toh, and David M Kingsley. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, 484(7392):55–61, April 2012.

[33] Jacob F Degner, Athma A Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J Gaffney, Joseph K Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–394, April 2013.

[34] Christopher D Brown, Lara M Mangravite, and Barbara E Engelhardt. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genetics*, 9(8):e1003649, 2013.

[35] Bas van Steensel and Job Dekker. Genomics tools for unraveling chromosome architecture. *Nature biotechnology*, 28(10):1089–1095, October 2010.

[36] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing

chromosome conformation. *Science (New York, NY)*, 295(5558):1306–1311, February 2002.

[37] Nele Gheldof, Marion Leleu, Daan Noordermeer, Jacques Rougemont, and Alexandre Reymond. Detecting long-range chromatin interactions using the chromosome conformation capture sequencing (4C-seq) method. *Methods in molecular biology (Clifton, N.J.)*, 786:211–225, 2012.

[38] Josée Dostie and Job Dekker. Mapping networks of physical interactions between genomic elements using 5C technology. *Nature protocols*, 2(4):988–1002, 2007.

[39] Nynke L van Berkum, Erez Lieberman-Aiden, Louise Williams, Maxim Imakaev, Andreas Gnirke, Leonid A Mirny, Job Dekker, and Eric S Lander. Hi-C: a method to study the three-dimensional architecture of genomes. *Journal of visualized experiments : JoVE*, (39), 2010.

[40] Melissa J Fullwood and Yijun Ruan. ChIP-based methods for the identification of long-range chromatin interactions. *Journal of cellular biochemistry*, 107(1):30–39, May 2009.

[41] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, Yiwen Chen, Gilberto DeSalvo, Charles Epstein, Katherine I Fisher-Aylor, Ghia Euskirchen, Mark Gerstein, Jason Gertz, Alexander J Hartemink, Michael M Hoffman, Vishwanath R Iyer, Youngsook L Jung, Subhradip Karmakar, Manolis Kellis, Peter V Kharchenko, Qunhua Li, Tao Liu, X Shirley Liu, Lijia Ma, Aleksandar Milosavljevic, Richard M Myers, Peter J Park, Michael J Pazin, Marc D Perry, Debasish Raha, Timothy E Reddy, Joel Rozowsky, Noam Shoresh, Arend Sidow, Matthew Slattery, John A Stamatoyannopoulos, Michael Y Tolstorukov, Kevin P White, Simon Xi, Peggy J Farnham, Jason D Lieb, Barbara J Wold, and Michael Snyder. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22(9):1813–1831, September 2012.

[42] U.S. Department of Energy. File:`chip_seq_illustration_final-hr.jpg` — brookhaven national lab newsroom, 2014. URL `\url{http://www.bnl.gov/bnlweb/pubaf/pr/photos/2011/11/chip_seq_illustration_final-`

hr.jpg}. [http://www.bnl.gov/bnlweb/pubaf/pr/photos/2011/11/chip_seq_illustration_final-hr.jpg; accessed 17-December-2014].

[43] Ho Sung Rhee and B Franklin Pugh. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell*, 147(6): 1408–1419, December 2011.

[44] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, January 2002.

[45] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, 41(Database issue):D991–5, January 2013.

[46] Wikipedia. File:chip-exo process diagram.pdf — wikipedia, the free encyclopedia, 2014. URL \url{http://en.wikipedia.org/wiki/File:ChIP-exo_process_diagram.pdf}. [http://en.wikipedia.org/wiki/File:ChIP-exo_process_diagram.pdf; accessed 17-December-2014].

[47] Benjamin L Kidder, Gangqing Hu, and Keji Zhao. ChIP-Seq: technical considerations for obtaining high-quality data. *Nature immunology*, 12(10):918–922, October 2011.

[48] R Ogata and W Gilbert. Contacts between the lac repressor and the thymines in the lac operator. *Proceedings of the National Academy of Sciences of the United States of America*, 74(11):4973–4976, November 1977.

[49] L Johnsrud. Contacts between Escherichia coli RNA polymerase and a lac operon promoter. *Proceedings of the National Academy of Sciences of the United States of America*, 75(11):5314–5318, November 1978.

[50] D J Galas and A Schmitz. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic acids research*, 5(9): 3157–3170, September 1978.

[51] Hashem Koohy, Thomas A Down, and Tim J Hubbard. Chromatin Accessibility Data Sets Show Bias Due to Sequence Specificity of the DNase I Enzyme. *PloS one*, 8(7):e69853, 2013.

[52] Housheng Hansen He, Clifford A Meyer, Sheng'en Shawn Hu, Mei-Wei Chen, Chongzhi Zang, Yin Liu, Prakash K Rao, Teng Fei, Han Xu, Henry Long, X Shirley Liu, and Myles Brown. refined dnase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature Methods*, pages 1–8, December 2013.

[53] Myong-Hee Sung, Michael J Guertin, Songjoon Baek, and Gordon L Hager. DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence. *Molecular cell*, 56(2):275–285, October 2014.

[54] Galip Gürkan Yardmc, Christopher L Frank, Gregory E Crawford, and Uwe Ohler. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic acids research*, 42(19):11865–11878, October 2014.

[55] Wikimedia Commons. File:`Courtney_2008.jpg`— wikimedia commons, the free media repository, 2014. URL `\url{http://commons.wikimedia.org/wiki/File:Courtney_2008.jpg}`. [`http://commons.wikimedia.org/wiki/File:Courtney_2008.jpg`; accessed 17-December-2014].

[56] P R Mueller and B Wold. In vivo footprinting of a muscle specific enhancer by ligation mediated PCR. *Science (New York, NY)*, 246(4931):780–786, November 1989.

[57] Peter J Sabo, Richard Humbert, Michael Hawrylycz, James C Wallace, Michael O Dorschner, Michael McArthur, and John A Stamatoyannopoulos. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proceedings of the National Academy of Sciences of the United States of America*, 101(13):4537–4542, March 2004.

[58] G E Crawford. From the Cover: Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proceedings of the National Academy of Sciences*, 101(4):992–997, January 2004.

[59] Lingyun Song and Gregory E Crawford. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor protocols*, 2010(2):pdb.prot5384, February 2010.

[60] Gregory E Crawford, Sean Davis, Peter C Scacheri, Gabriel Renaud, Mohamad J Halawi, Michael R Erdos, Roland Green, Paul S Meltzer, Tyra G

Wolfsberg, and Francis S Collins. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature Methods*, 3(7):503–509, July 2006.

[61] Jay R Hesselberth, Xiaoyu Chen, Zhihong Zhang, Peter J Sabo, Richard Sandstrom, Alex P Reynolds, Robert E Thurman, Shane Neph, Michael S Kuehn, William S Noble, Stanley Fields, and John A Stamatoyannopoulos. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, 6(4):283–289, April 2009.

[62] Weihua Zeng and Ali Mortazavi. Technical considerations for functional sequencing assays. *Nature immunology*, 13(9):802–807, September 2012.

[63] Peter J Sabo, Michael S Kuehn, Robert Thurman, Brett E Johnson, Ericka M Johnson, Hua Cao, Man Yu, Elizabeth Rosenzweig, Jeff Goldy, Andrew Haydock, Molly Weaver, Anthony Shafer, Kristin Lee, Fidencio Neri, Richard Humbert, Michael A Singer, Todd A Richmond, Michael O Dorschner, Michael McArthur, Michael Hawrylycz, Roland D Green, Patrick A Navas, William S Noble, and John A Stamatoyannopoulos. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature Methods*, 3(7):511–518, July 2006.

[64] Jeff Vierstra, Hao Wang, Sam John, Richard Sandstrom, and John A Stamatoyannopoulos. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nature Methods*, 11(1):66–72, November 2013.

[65] Teemu D Laajala, Sunil Raghav, Soile Tuomela, Riitta Lahesmaa, Tero Aittokallio, and Laura L Elo. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC genomics*, 10:618, 2009.

[66] Shane Neph, Jeff Vierstra, Andrew B Stergachis, Alex P Reynolds, Eric Haugen, Benjamin Vernot, Robert E Thurman, Sam John, Richard Sandstrom, Audra K Johnson, Matthew T Maurano, Richard Humbert, Eric Rynes, Hao Wang, Shinny Vong, Kristen Lee, Daniel Bates, Morgan Diegel, Vaughn Roach, Douglas Dunn, Jun Neri, Anthony Schafer, R Scott Hansen, Tanya Kutyavin, Erika Giste, Molly Weaver, Theresa Canfield, Peter Sabo, Miaohua Zhang, Gayathri Balasundaram, Rachel Byron, Michael J MacCoss, Joshua M Akey, M A Bender, Mark Groudine, Rajinder Kaul, and John A Stamatoy-

annopoulos. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, September 2012.

[67] Roger Pique-Regi, Jacob F Degner, Athma A Pai, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3):447–455, March 2011.

[68] Alan P Boyle, Lingyun Song, Bum-Kyu Lee, Darin London, Damian Keefe, Ewan Birney, Vishwanath R Iyer, Gregory E Crawford, and Terrence S Furey. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, 21(3):456–464, March 2011.

[69] Dan Graur, Yichen Zheng, Nicholas Price, Ricardo B R Azevedo, Rebecca A Zufall, and Eran Elhaik. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome biology and evolution*, 5(3):578–590, 2013.

[70] G D Stormo, T D Schneider, L Gold, and A Ehrenfeucht. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic acids research*, 10(9):2997–3011, May 1982.

[71] Martin Tompa, Nan Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W James Kent, Vsevolod J Makeev, Andrei A Mironov, William Stafford Noble, Giulio Pavesi, Graziano Pesole, Mireille Régnier, Nicolas Simonis, Saurabh Sinha, Gert Thijs, Jacques van Helden, Mathias Vandenbogaert, Zhiping Weng, Christopher Workman, Chun Ye, and Zhou Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, 23(1):137–144, January 2005.

[72] V Matys, O V Kel-Margoulis, E Fricke, I Liebich, S Land, A Barre-Dirrie, I Reuter, D Chekmenev, M Krull, K Hornischer, N Voss, P Stegmaier, B Lewicki-Potapov, H Saxel, A E Kel, and E Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(Database issue):D108–10, January 2006.

[73] Anthony Mathelier, Xiaobei Zhao, Allen W Zhang, François Parcy, Rebecca Worsley-Hunt, David J Arenillas, Sorana Buchman, Chih-yu Chen, Alice Chou, Hans Ienasescu, Jonathan Lim, Casper Shyr, Ge Tan, Michelle Zhou,

Boris Lenhard, Albin Sandelin, and Wyeth W Wasserman. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic acids research*, 42(Database issue):D142–7, January 2014.

[74] A D Sharrocks. The ETS-domain transcription factor family. *Nature reviews. Molecular cell biology*, 2(11):827–837, November 2001.

[75] Jeff Vierstra, Eric Rynes, Richard Sandstrom, Miaohua Zhang, Theresa Canfield, R Scott Hansen, Sandra Stehling-Sun, Peter J Sabo, Rachel Byron, Richard Humbert, Robert E Thurman, Audra K Johnson, Shinny Vong, Kristen Lee, Daniel Bates, Fidencio Neri, Morgan Diegel, Erika Giste, Eric Haugen, Douglas Dunn, Matthew S Wilken, Steven Josefowicz, Robert Samstein, Kai-Hsin Chang, Evan E Eichler, Marella De Bruijn, Thomas A Reh, Arthur Skoultchi, Alexander Rudensky, Stuart H Orkin, Thalia Papayannopoulou, Piper M Treuting, Licia Selleri, Rajinder Kaul, Mark Groudine, M A Bender, and John A Stamatoyannopoulos. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science (New York, NY)*, 346 (6212):1007–1012, November 2014.

[76] Xiao-Yong Li, Sean Thomas, Peter J Sabo, Michael B Eisen, John A Stamatoyannopoulos, and Mark D Biggin. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome biology*, 12(4):R34, 2011.

[77] Wenli Zhang, Tao Zhang, Yufeng Wu, and Jiming Jiang. Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. *The Plant cell*, 24(7):2719–2731, July 2012.

[78] Yoichiro Shibata, Nathan C Sheffield, Olivier Fedrigo, Courtney C Babbitt, Matthew Wortham, Alok K Tewari, Darin London, Lingyun Song, Bum-Kyu Lee, Vishwanath R Iyer, Stephen C J Parker, Elliott H Margulies, Gregory A Wray, Terrence S Furey, and Gregory E Crawford. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. *PLoS Genetics*, 8(6): e1002789, June 2012.

[79] Hashem Koohy, Thomas A Down, Mikhail Spivakov, and Tim Hubbard. A

comparison of peak callers used for DNase-Seq data. *PloS one*, 9(5):e96303, 2014.

[80] Iros Barozzi, Pranami Bora, and Marco J Morelli. Comparative evaluation of DNase-seq footprint identification strategies. *Frontiers in genetics*, 5:278, 2014.

[81] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25, 2009.

[82] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, March 2012.

[83] Richard I Sherwood, Tatsunori Hashimoto, Charles W O'Donnell, Sophia Lewis, Amira A Barkal, John Peter van Hoff, Vivek Karun, Tommi Jaakkola, and David K Gifford. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature biotechnology*, 32(2):171–178, February 2014.

[84] Leighton J Core, André L Martins, Charles G Danko, Colin T Waters, Adam Siepel, and John T Lis. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature genetics*, November 2014.

[85] Anetta Ptasinska, Salam A Assi, Natalia Martinez-Soria, Maria Rosaria Imperato, Jason Piper, Pierre Cauchy, Anna Pickin, Sally R James, Maarten Hoogenkamp, Dan Williamson, Mengchu Wu, Daniel G Tenen, Sascha Ott, David R Westhead, Peter N Cockerill, Olaf Heidenreich, and Constanze Bonifer. Identification of a dynamic core transcriptional network in t(8;21) AML that regulates differentiation block and self-renewal. *Cell reports*, 8(6): 1974–1988, September 2014.

[86] S E Langabeer, H Walker, J R Rogers, A K Burnett, K Wheatley, D Swirsky, A H Goldstone, and D C Linch. Incidence of AML1/ETO fusion transcripts in patients entered into the MRC AML trials. MRC Adult Leukaemia Working Party. *British journal of haematology*, 99(4):925–928, December 1997.

[87] M Mitterbauer, R Kusec, I Schwarzinger, O A Haas, K Lechner, and U Jaeger. Comparison of karyotype analysis and RT-PCR for AML1/ETO in 204 un-

selected patients with AML. *Annals of hematology*, 76(3-4):139–143, March 1998.

[88] T Okuda, J van Deursen, S W Hiebert, G Grosveld, and J R Downing. AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis. *Cell*, 84(2):321–330, January 1996.

[89] A Ptasinska, S A Assi, D Mannari, S R James, D Williamson, J Dunne, M Hoogenkamp, M Wu, M Care, H McNeill, P Cauchy, M Cullen, R M Tooze, D G Tenen, B D Young, P N Cockerill, D R Westhead, O Heidenreich, and C Bonifer. Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding. *Leukemia*, 26(8):1829–1841, August 2012.

[90] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, Peggy J Farnham, Martin Hirst, Eric S Lander, Tarjei S Mikkelsen, and James A Thomson. The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, 28(10):1045–1048, October 2010.

[91] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, August 2009.

[92] Pedro Madrigal and Paweł Krajewski. Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. *Frontiers in genetics*, 3:230, 2012.

[93] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, Viren Amin, John W Whitaker, Matthew D Schultz, Lucas D Ward, Abhishek Sarkar, Gerald Quon, Richard S Sandstrom, Matthew L Eaton, Yi-Chieh Wu, Andreas R Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R Alan Harris, Noam Shoresh, Charles B Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J Mungall, Richard Moore, Eric Chuah, Angela Tam, Theresa K

Canfield, R Scott Hansen, Rajinder Kaul, Peter J Sabo, Mukul S Bansal, Annaick Carles, Jesse R Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca Lowdon, GiNell Elliott, Tim R Mercer, Shane J Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C Sallari, Kyle T Siebenthall, Nicholas A Sinnott-Armstrong, Michael Stevens, Robert E Thurman, Jie Wu, Bo Zhang, Xin Zhou, Arthur E Beaudet, Laurie A Boyer, Philip L De Jager, Peggy J Farnham, Susan J Fisher, David Haussler, Steven J M Jones, Wei Li, Marco A Marra, Michael T McManus, Shamil Sunyaev, James A Thomson, Thea D Tlsty, Li-Huei Tsai, Wei Wang, Robert A Waterland, Michael Q Zhang, Lisa H Chadwick, Bradley E Bernstein, Joseph F Costello, Joseph R Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, February 2015.

[94] Gregor D Gilfillan, Timothy Hughes, Ying Sheng, Hanne S Hjorthaug, Tobias Straub, Kristina Gervin, Jennifer R Harris, Dag E Undlien, and Robert Lyle. Limitations and possibilities of low cell number ChIP-seq. *BMC genomics*, 13:645, 2012.

[95] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, December 2013.

[96] S A Bustin and I A McKay. Transcription factors: targets for new designer drugs. *British journal of biomedical science*, 51(2):147–157, June 1994.

[97] Michalis V Karamouzis and Athanasios G Papavassiliou. Transcription factor networks as targets for therapeutic intervention of cancer: the breast cancer paradigm. *Molecular medicine (Cambridge, Mass.)*, 17(11-12):1133–1136, 2011.

[98] Panagiotis A Konstantinopoulos and Athanasios G Papavassiliou. Seeing the future of cancer-associated transcription factor drug targets. *JAMA*, 305(22): 2349–2350, June 2011.

[99] T Liu and R B Altman. Identifying druggable targets by protein microenvironments matching: application to transcription factors. *CPT: pharmacometrics & systems pharmacology*, 3:e93, 2014.

Appendix A

Identification of a dynamic core transcriptional network in t(8;21) AML regulating differentiation block and self-renewal

OPEN ACCESS
CellPress

Cell Reports
Article

Identification of a Dynamic Core Transcriptional Network in t(8;21) AML that Regulates Differentiation Block and Self-Renewal

Anetta Ptasinska,[1,7] Salam A. Assi,[1,3,7] Natalia Martinez-Soria,[4] Maria Rosaria Imperato,[1] Jason Piper,[2] Pierre Cauchy,[1] Anna Pickin,[1] Sally R. James,[6] Maarten Hoogenkamp,[1] Dan Williamson,[4] Mengchu Wu,[5] Daniel G. Tenen,[5] Sascha Ott,[2] David R. Westhead,[3] Peter N. Cockerill,[1] Olaf Heidenreich,[4,*] and Constanze Bonifer[1,*]

[1]School of Cancer Sciences, College of Medicine and Dentistry, University of Birmingham, Birmingham B15 2TT, UK
[2]Warwick Systems Biology Centre, University of Warwick, Coventry CV4 7AL, UK
[3]School of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK
[4]Northern Institute for Cancer Research, University of Newcastle, Newcastle upon Tyne NE2 4HH, UK
[5]Cancer Science Institute, National University of Singapore, Republic of Singapore, Singapore 117456, Singapore
[6]Section of Experimental Haematology, Leeds Institute for Molecular Medicine, University of Leeds, Leeds LS2 9JT, UK
[7]Co-first author
*Correspondence: olaf.heidenreich@ncl.ac.uk (O.H.), c.bonifer@bham.ac.uk (C.B.)
http://dx.doi.org/10.1016/j.celrep.2014.08.024

SUMMARY

Oncogenic transcription factors such as RUNX1/ETO, which is generated by the chromosomal translocation t(8;21), subvert normal blood cell development by impairing differentiation and driving malignant self-renewal. Here, we use digital footprinting and chromatin immunoprecipitation sequencing (ChIP-seq) to identify the core RUNX1/ETO-responsive transcriptional network of t(8;21) cells. We show that the transcriptional program underlying leukemic propagation is regulated by a dynamic equilibrium between RUNX1/ETO and RUNX1 complexes, which bind to identical DNA sites in a mutually exclusive fashion. Perturbation of this equilibrium in t(8;21) cells by RUNX1/ETO depletion leads to a global redistribution of transcription factor complexes within preexisting open chromatin, resulting in the formation of a transcriptional network that drives myeloid differentiation. Our work demonstrates on a genome-wide level that the extent of impaired myeloid differentiation in t(8;21) is controlled by the dynamic balance between RUNX1/ETO and RUNX1 activities through the repression of transcription factors that drive differentiation.

INTRODUCTION

Lineage-specific cell differentiation is controlled by the establishment of specific gene-expression patterns in normal cells, and interference with this process underpins oncogenesis. Hematopoiesis is one of the best-understood developmental pathways and involves dynamic alterations in transcriptional programs, which regulate progression along the differentiation hierarchy (Pimanda and Göttgens, 2010). Individual cellular differentiation states are defined by transcriptional networks composed of combinations of transcription factors that bind to specific sets of *cis*-regulatory elements (Davidson, 2010). Therefore, experimental analysis of the binding activities of multiple factors has served as a means of identifying crucial regulators for a specific cell type (DeVilbiss et al., 2014; Tijssen et al., 2011). However, normal differentiation is impaired in cancers, leading cells to adopt a new malignant identity. Unique insights into processes that control development toward both normal and perturbed differentiation states can be gained from a detailed examination of the mechanisms utilized by leukemic transcription factors such as PML/RARA, MLL fusion proteins, and RUNX1/ETO. These factors reprogram the epigenome and thereby block the hierarchical succession of normal transcriptional networks.

Leukemias are characterized by good experimental accessibility and, compared with many carcinomas, relatively high genetic stability, which makes them very amenable to investigations of general as well as specific mechanisms of oncogenesis. Acute myeloid leukemia (AML) is the second most common leukemia and is a heterogeneous disease with impaired myeloid differentiation (Valk et al., 2004). The hallmarks of AML are multiple somatic mutations, including genetic rearrangements that affect signal transduction and gene expression. This includes mutations in genes encoding DNA methylases, chromatin modifiers, and transcription factors. Many such mutations affect transcription factors that are crucial for the development of hematopoietic stem cells or for terminal myeloid differentiation, such as RUNX1 and C/EBPα, respectively (Gaidzik et al., 2011; Michaud et al., 2002; Pabst et al., 2001b; Snaddon et al., 2003). However, the molecular details of how such mutant transcription factors cause alterations of the epigenome are still insufficiently understood. In addition, so far no experiments have defined the core transcriptional network of a specific type of AML and dissected the role of mutated transcription factors within this network.

One of the best-characterized chromosomal rearrangements found in AML is the t(8;21) translocation, which accounts for approximately 10% of all AMLs. This translocation fuses the DNA-binding domain of the hematopoietic master regulator RUNX1 to almost the entire ETO protein, which is an adaptor protein for histone deacetylase (HDAC) complexes (Miyoshi et al., 1993). The resulting RUNX1/ETO fusion protein lacks the transactivation domain of RUNX1, resulting in major differences in the biological activities of the two proteins. RUNX1 normally recruits transcriptional activators and binds to DNA as a heterodimer with core-binding factor β (CBFβ). The RUNX1/ETO fusion protein also interacts with CBFβ but functions as a RUNX1/ETO tetramer (Liu et al., 2006), and like ETO itself, it also interacts with NCOR and SIN3A corepressors (Amann et al., 2001). Consequently, this chromosomal rearrangement converts a transcriptional activator into a repressor. However, there is evidence that RUNX1 also interacts with HDACs via SIN3A and can act as a repressor (Reed-Inderbitzin et al., 2006; Taniuchi et al., 2002). Proteomic and chromatin immunoprecipitation (ChIP) analyses in t(8;21) cell lines have demonstrated the association of RUNX1/ETO with multiple hematopoietic regulators known to be involved in the regulation of hematopoietic stem cell genes (Wilson et al., 2010). The RUNX1/ETO complex consists of the E box binding transcription factors HEB and LYL1 and the bridging factors LMO2 and LDB1. In chromatin, this complex interacts with the ETS family members FLI1 and ERG, and these interactions are required for the stability of the complex and its leukemogenicity (Martens et al., 2012; Sun et al., 2013).

Genome-wide analyses in t(8;21) cell lines and in patients via ChIP sequencing (ChIP-seq) identified thousands of RUNX1/ETO-binding sites (Ben-Ami et al., 2013; Martens et al., 2012; Ptasinska et al., 2012; Saeed et al., 2012), but the role of specific binding sites within the AML-specific transcriptional network is unclear. All t(8;21) AML cells retain an intact copy of RUNX1, which is required for cell survival—a feature that has also been observed in other CBF leukemias (Ben-Ami et al., 2013; Goyama et al., 2013). RUNX1 and RUNX1/ETO each drive the expression of alternate subsets of genes (Ben-Ami et al., 2013). However, 60% of the RUNX1/ETO sites are shared with RUNX1 (Ptasinska et al., 2012), and whether there is a direct dynamic competition between RUNX1/ETO and RUNX1 for the same genomic sites remains to be investigated.

The differentiation of t(8;21) cells is blocked at an early myeloid progenitor stage and so far the core transcriptional program underlying this block has been elusive. Changes in RUNX1/ETO expression in t(8;21) AML cells are associated with both up- and downregulated genes, and individual RUNX1/ETO-bound genomic sites recruit both histone acetyltransferases (HATs) and HDACs (Follows et al., 2003; Ptasinska et al., 2012; Sun et al., 2013; Wang et al., 2011). However, we previously showed that the genome-wide loss of RUNX1/ETO binding correlates with increased histone H3 lysine 9 (H3K9) acetylation (Ptasinska et al., 2012). In addition, RUNX1/ETO depletion is associated with the upregulation of C/EBPα, a driver of myeloid and, in particular, granulocytic differentiation (Zhang et al., 1997). Moreover, RUNX1/ETO has been shown to sequester C/EBPα from its murine promoter, thereby interfering with C/EBPα expression (Pabst et al., 2001a). RUNX1/ETO knockdown causes release of the differentiation block, resulting in a gene-expression pattern that resembles that of granulocytes and monocytes (Ptasinska et al., 2012). Taken together, these results suggest that RUNX1/ETO-mediated reprogramming of the epigenome involves a complex and so far unexplored interplay of different transcription-factor and chromatin-modifying cofactor activities. To date, we have gained little insight into the nature of this reprogrammed network and the sequential order of factors required to restore normal myeloid cell functions.

In this study, we addressed these issues by investigating the dynamic changes in global transcription-factor-binding patterns that occur following depletion of RUNX1/ETO. To that end, we combined ChIP-seq for multiple factors, DNaseI footprinting, and transcriptome analysis to identify the core transcriptional network of t(8;21) AML cells, and then characterized changes in these networks upon RUNX1/ETO knockdown. These analyses revealed a dynamic equilibrium between RUNX1/ETO and RUNX1 complexes competing for identical genomic sites. Results from sequential ChIP (re-ChIP) show that the two complexes have similar accessory-factor compositions but differ in their preference for the recruitment of coactivators and corepressors. Using a digital DNaseI footprinting approach, we found that both t(8;21)-positive cell lines (Kasumi-1 and SKNO-1) and patient-derived primary AML cells with the t(8;21) translocation (patient cells) share the same pattern of binding-site occupancy. Within this core transcriptional network, RUNX1/ETO-bound loci are predominantly associated with transcriptional repression. Furthermore, loss of RUNX1/ETO establishes a differentiation-associated transcriptional network dominated by de novo binding of C/EBPα resulting from the upregulation of *CEBPA* gene expression. Our results demonstrate that the block in myeloid differentiation in t(8;21) AML results from the dynamic interference of RUNX1/ETO with *cis*-regulatory elements that normally are destined to change transcription-factor assemblies during myeloid differentiation, notably those that increase binding of RUNX1 and C/EBPα.

RESULTS

Transcription-Factor Occupancy Patterns Are Highly Comparable between t(8;21) Cell Lines and Patient Cells

To define the RUNX1/ETO-responsive core transcriptional network and monitor dynamic changes associated with alterations in RUNX1/ETO status, we utilized Kasumi-1 cells, which represent a well characterized and widely used model system for t(8;21) AML (Ben-Ami et al., 2013; Martens et al., 2012; Ptasinska et al., 2012; Sun et al., 2013). We measured the binding of multiple transcription factors in these cells using genome-wide ChIP-seq and performed perturbation experiments by transiently knocking down RUNX1/ETO expression. We then monitored the consequences using ChIP-seq and RNA sequencing (RNA-seq) analyses (Heidenreich et al., 2003; Ptasinska et al., 2012; Table S1). We used antibodies against RUNX1, the ETO moiety of RUNX1/ETO, LMO2 as a member of the RUNX1/ETO complex, RNA-Polymerase II, and acetylated histone H3 for ChIP. To obtain a more complete picture of the composition of RUNX1 and RUNX1/ETO-associated transcription-factor complexes

without RUNX1/ETO knockdown, we also analyzed publicly available data for the E box protein HEB (Martens et al., 2012; Ptasinska et al., 2012). In order to follow additional alterations in the epigenome after RUNX1/ETO knockdown, we also measured the binding of PU.1 and C/EBPα, which are both required for myeloid differentiation (Scott et al., 1994; Zhang et al., 1997). We identified high-confidence transcription-factor binding-site peaks by integrating ChIP data with DNaseI-seq data before and after RUNX1/ETO depletion, and considered only those peaks that were located within DNaseI hypersensitive sites (DHSs).

RUNX1/ETO exists as a complex with other transcription factors (Sun et al., 2013). Consistent with these findings, we observed a colocalization of RUNX1/ETO, RUNX1, HEB, LMO2, C/EBPα, and/or PU.1 binding at many DHSs in Kasumi-1 cells, as exemplified by the *LMO2* locus (Figure 1A). Closer examination of the genome-wide occupancy patterns of LMO2 and HEB revealed that a substantial overlap existed among LMO2, HEB, and RUNX1/ETO binding sites (Figure S1A). Although there was some overlap with the other factors, the PU.1 and C/EBPα binding sites did not closely cluster as a group with those for the RUNX1/ETO complexes in Kasumi-1.

We next sought to determine whether the RUNX1/ETO and RUNX1 binding patterns identified in Kasumi-1 cells were shared with patient cells. First, we performed a DHS analysis on patient cells and normal CD34+ hematopoietic stem and precursor cells (CD34+ cells) derived from the peripheral blood of healthy donors. This fraction is enriched for stem and multipotent progenitor cells. DHS mapping was complemented by RUNX1/ETO and RUNX1 ChIP analysis. However, the large quantity of material required for this approach precluded analysis of patient cells. Therefore, to determine which subsets of DHSs from patient cells overlap with sites that recruit RUNX1 and RUNX1/ETO in the cell line and in CD34+ cells, we first generated a scatter diagram of the joint DHS signal of patient cells (Ptasinska et al., 2012) compared with normal CD34+ cells (Figure S1B). We then projected the genomic coordinates from the RUNX1/ETO and RUNX1 ChIP experiments onto these sequences. These diagrams clearly show that the RUNX1- and RUNX1/ETO-bound sequences from Kasumi-1 cells projected onto the DHS peaks from patient cells, whereas RUNX1-bound sequences from CD34+ cells projected onto the DHS peaks from the CD34+ cells.

To further confirm the similarity between t(8;21) cell lines and patient cells, and to test whether we could overcome the need to conduct multiple ChIP-seq experiments, we generated additional higher-read-depth DNaseI data from two t(8;21) patients and developed a digital footprinting algorithm (Wellington). This high-resolution approach takes the chromatin structure surrounding transcription-factor motifs that are protected from DNaseI digestion into account and thus evaluates the genome-wide transcription-factor occupancy with high accuracy (Piper et al., 2013). DNaseI footprinting data obtained from one t(8;21) patient were compared with ChIP data for regions bound by RUNX1/ETO, RUNX1, HEB, and LMO2 in Kasumi-1 cells (13,584 peaks in total). This comparison demonstrated a high concordance between transcription-factor binding in Kasumi-1 cells and motif occupancy in patient cells, as defined by preferential protection against DNaseI digestion (Figure S1C). This is exemplified by the DNaseI footprints found at the *NFE2* locus (Figure 1B, gray areas), which in both patient samples reflect the pattern of binding of RUNX1/ETO, HEB, LMO2, PU.1, and RUNX1 in Kasumi-1 cells. These sites also form a DHS in normal CD34+ cells and are bound by RUNX1 in these cells, as determined by ChIP (Figure 1B, top).

In contrast to RUNX1, which interacts with a multiplicity of factors in different cell types (Scheitz and Tumbar, 2013; van Riel et al., 2012), RUNX1/ETO preferentially binds to DNA elements containing RUNX, ETS, and E box motifs, thus reflecting the composition of the RUNX1/ETO complex (Sun et al., 2013). To examine whether our footprinting analysis was able to confirm this preference of colocalizing motifs in patient cells, we conducted an unbiased pairwise clustering analysis of footprinted motifs in regions bound by RUNX1/ETO. This analysis demonstrated that motifs bound by RUNX1/ETO in Kasumi-1 cells strongly clustered with ETS (PU.1 and ERG) and E box (SCL, LYL, and HEB) motifs that are footprinted in patient cells (Figure 1C). We found a similar clustering pattern using sequences from the Kasumi-1 ChIP-seq experiments (Figure S1D), although it was less defined due to the larger peak sizes in this experimental context. In conclusion, RUNX1/ETO-positive Kasumi-1 cells show similar transcription-factor motif occupancy patterns, confirming that at this level of accuracy, digital footprinting provides a viable method for investigating transcription-factor binding-site occupancy and preferential interaction in patient cells.

RUNX1/ETO and RUNX1-Containing Complexes Compete for the Same Genomic Sites

We previously showed that more than 60% of RUNX1/ETO binding sites are shared with RUNX1 in the bulk population of cells (Ptasinska et al., 2012), with many of the footprinted sites containing multiple TGYGGT RUNX1-binding motifs (e.g., Figure 1B). Therefore, we conducted re-ChIP experiments in Kasumi-1 cells to test at known RUNX1/ETO binding sites whether the two factors co-occupy single sites or whether binding is mutually exclusive at such sites. In addition, we examined which other factors were shared between RUNX1 and RUNX1/ETO complexes. RUNX1 and RUNX1/ETO both colocalize with LMO2, HEB, and LYL1 in the Kasumi-1 cell population (Figures 1A, 2A, and S2A). However, binding of RUNX1 and RUNX1/ETO to their target sites was mutually exclusive, even at elements containing multiple RUNX motifs, such as the *NFE2* locus (Figures 1B, 2B, 2C, and S2B).

Both RUNX1/ETO and RUNX1 have been shown to interact with HDACs and the HAT p300 (also known as EP300) (Amann et al., 2001; Kitabayashi et al., 1998; Levanon et al., 1998; Reed-Inderbitzin et al., 2006; Wang et al., 2011). Using paralled re-ChIP experiments, we show that RUNX1-bound elements had a preference for binding the coactivator p300, whereas RUNX1/ETO-occupied elements preferentially bound the corepressor HDAC2 (Figures 2D–2F). We further confirmed this preferential binding and the strong association between RUNX1 and p300 by performing manual ChIP and ChIP-sequencing experiments after knockdown of RUNX1/ETO (Figure 3). These experiments demonstrated (1) that the loss of RUNX1/ETO binding led to an

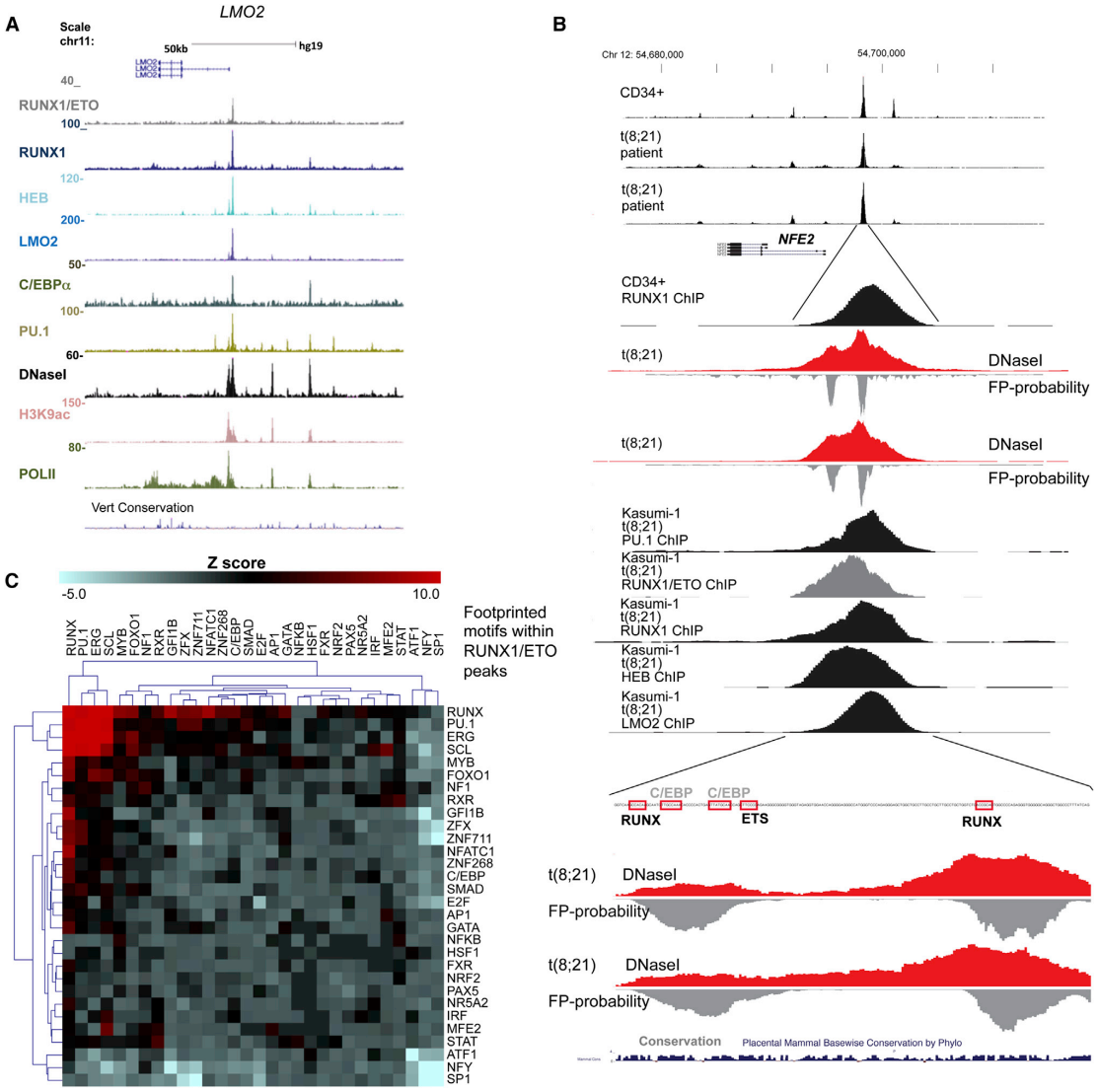


Figure 1. Transcription-Factor Occupancy Patterns Are Similar between RUNX1/ETO-Expressing Cell Lines and Patient Cells

(A) UCSC genome browser screenshot showing the binding patterns of RUNX1/ETO, RUNX1, HEB, LMO2, C/EBPα, PU.1, DHS, H3K9Ac, and RNA-Polymerase II (POLII), as well as input reads and conservation among vertebrates at the *LMO2* locus as aligned reads.

(B) UCSC genome browser screenshot of ChIP-seq and DHS data aligned with digital footprints at the *NFE2* locus within a DHS shared between two t(8;21) patients and purified normal CD34+ cells (top). It also shows the binding pattern of RUNX1 in CD34+ cells and RUNX1/ETO, RUNX1, HEB, LMO2, and PU.1 in Kasumi-1 cells as determined by ChIP. Footprint probabilities as calculated by Wellington are indicated as gray columns below the lines. The bottom indicates the location of occupied RUNX, ETS, and C/EBP motifs.

(C) Occupied RUNX, E box, and ETS motifs in patient cells cluster within DHS sites that colocalize with RUNX1/ETO binding in Kasumi-1 cells. The heatmap shows hierarchical clustering of footprinted motif co-occurrences by *Z* score within RUNX1/ETO peaks, indicating transcription factor co-occupancy. Footprint probabilities within RUNX1/ETO-bound peaks were calculated using DNaseI-seq data from t(8;21) patient 1. The motif search was done within RUNX1/ETO footprint coordinates. Red and blue colors indicate statistically over- and underrepresented motif co-occurrences, respectively. For a more detailed explanation, see the legend of Figure S1 and the Supplemental Experimental Procedures.

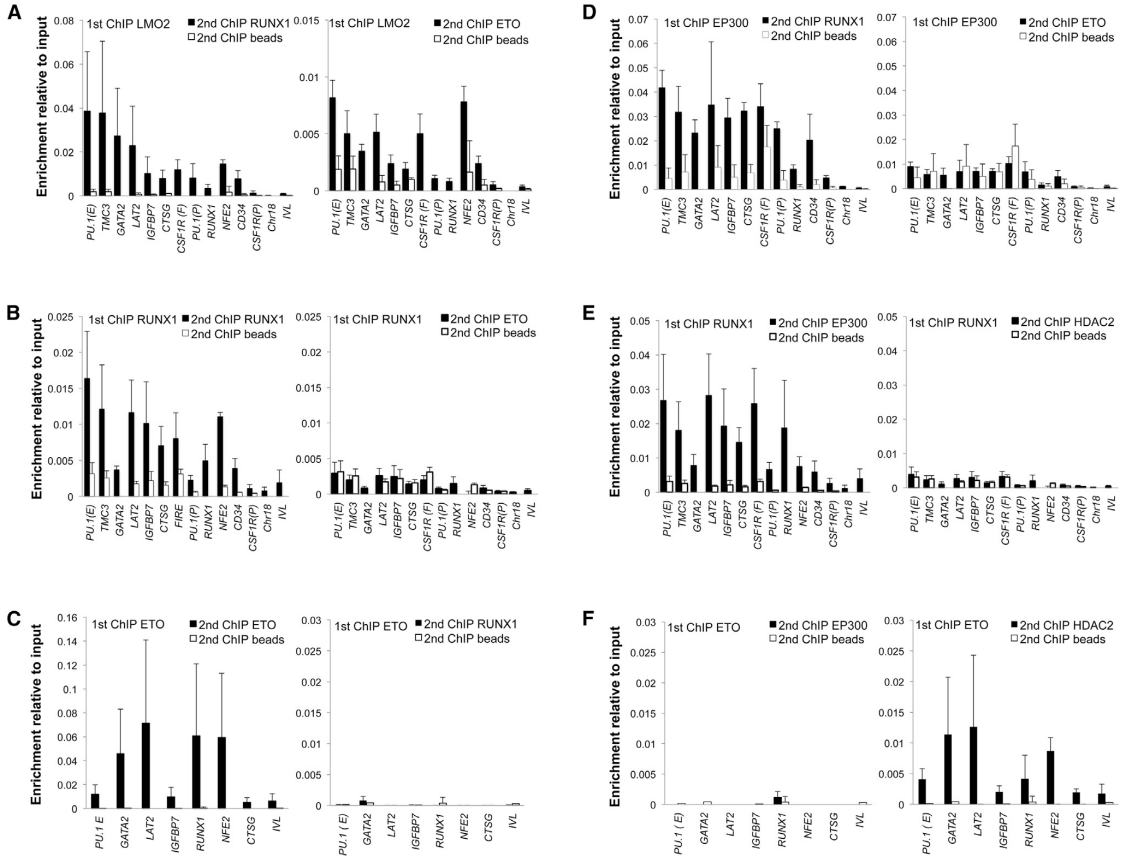


Figure 2. RUNX1 and RUNX1/ETO Complexes Differentially Interact with Coactivator and Corepressor Complexes, and Binding to the Same Sites Is Mutually Exclusive

(A–E) Multiple RUNX1/ETO binding sequences and control sequences (*IVL*, Chr18) were selected and validated for factor binding by a first round of ChIP followed by a second round with a different antibody or with just beads as indicated. All of the chosen binding sites contain several RUNX1 motifs (data not shown).

(A) LMO2 associates with both RUNX1 and RUNX1/ETO.

(B and C) RUNX1 and RUNX1/ETO binding is mutually exclusive. Control ChIPs were performed with the same antibody.

(D) EP300 associates with RUNX1, but not RUNX1/ETO.

(E and F) RUNX1 preferentially binds p300, whereas RUNX1/ETO preferentially associates with HDAC2. For additional amplicons, see Figure S2B. qPCR data represent the mean ± SD of at least three independent experiments.

increase in RUNX1 binding at the same sites, and (2) there was an increased recruitment of p300 without a concomitant increase in the expression of these factors (Figures 3 and S3A), providing an explanation for the increased histone H3 lysine 9 acetylation at such sites that we observed previously (Figure S3B; Ptasinska et al., 2012). In contrast, knockdown of RUNX1/ETO led to a reduction of HDAC2 binding to these target sites (Figure 3C). Taken together, these data show that RUNX1/ETO and RUNX1 (1) compete for the same genomic sites and (2) colocalize with the same transcription factors but have distinct preferences for histone-modifying cofactors, with RUNX1 associated complexes preferring to interact with p300 and RUNX1/ETO complexes preferring to recruit HDACs, including HDAC2.

The Core Transcriptional Network Bound by RUNX1/ETO Is Predominantly Associated with Repressed Genes

We next analyzed our ChIP-seq data sets to identify the core transcriptional network that characterizes the cellular identity of t(8;21) cells by determining overrepresented combinatorial binding patterns for the transcription factors RUNX1/ETO, C/EBPα, HEB, LMO2, PU.1, and RUNX1 (Tijssen et al., 2011). ChIP sequences in RUNX1/ETO-positive cells were enriched for just 11 of the 63 possible different binding patterns, which included six significantly enriched combinatorial patterns containing RUNX1/ETO and five patterns that did not (Figure 4A, marked by asterisks). Two possible binding patterns (111010 and 110011) were not observed. We then associated such

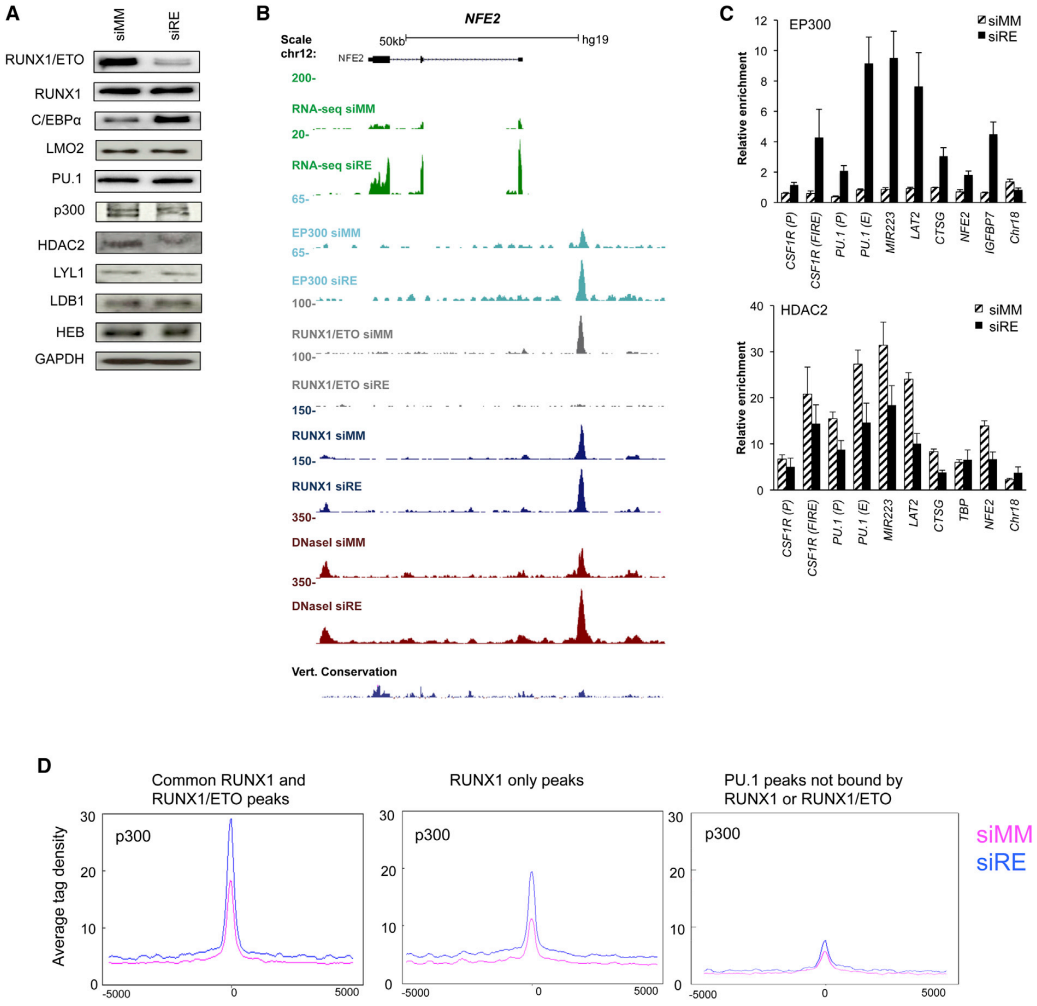


Figure 3. Dynamic Alterations in Cofactor Binding upon RUNX1/ETO Knockdown

(A) Western blot detecting RUNX1/ETO, RUNX1, C/EBPα, LMO2, PU.1, p300, HDAC2, LYL1, LDB1, and HEB protein in Kasumi-1 cells treated for 48 hr with mismatch control siRNA (siMM) and with RUNX1/ETO siRNA (siRE). GAPDH served as the loading control.

(B) UCSC genome browser screenshot of the *NFE2* locus showing changes in the RNA expression and binding pattern of p300, RUNX1/ETO (R/E), RUNX1, and DHS upon RUNX/ETO knockdown in Kasumi-1 cells.

(C) Increase of p300 binding and decrease of HDAC2 binding upon RUNX1/ETO knockdown.

(D) Global changes of p300 binding peaks shared between RUNX1/ETO and RUNX, peaks exclusively bound by RUNX1, and PU.1 peaks not associated with RUNX1/ETO or RUNX1 binding. qPCR data represent the mean ± SD of three to five independent experiments. For other control analyses, see Figure S3B.

elements with the nearest genes and performed a gene set enrichment analysis (GSEA) using gene-expression data sets derived from a time course of RUNX1/ETO knockdown in two different t(8;21) cell lines (Figures S4A and S4B; Ptasinska et al., 2012). In addition, we compared these gene signatures with a RNA-seq-based gene-expression data set derived from a 4-day RUNX1/ETO knockdown in Kasumi-1 cells (Figures 4B and S4C). This analysis demonstrated that all overrepresented RUNX1/ETO-containing binding patterns were associated with the upregulation of gene expression upon knockdown (Figure 4B, red asterisks), whereas loci that do not bind RUNX1/ETO were enriched in genes that were downregulated after RUNX1/ETO knockdown (green asterisks). The very same genes behaved similarly when assayed after knockdown of RUNX1/ETO in patient cells, confirming the similarity between cell lines and primary cells (Figure 4C).

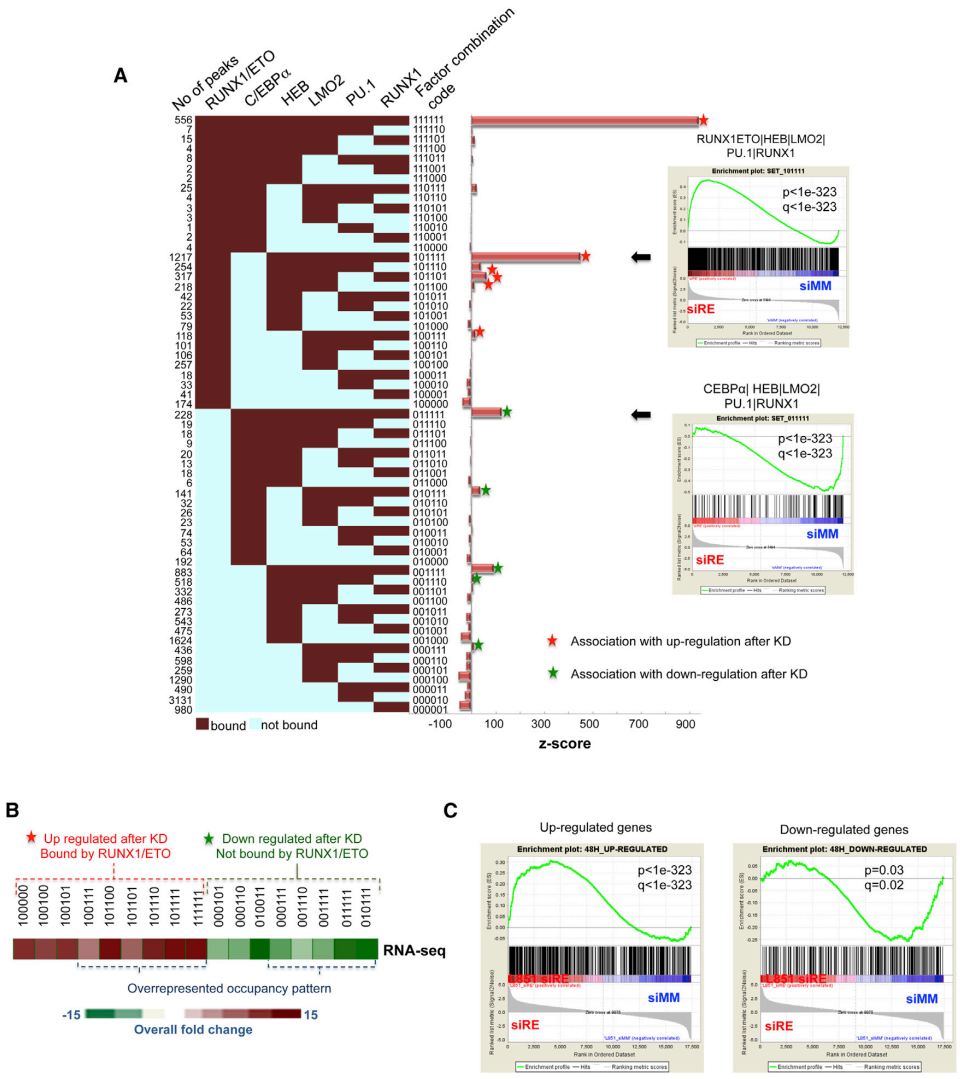


Figure 4. Specific Transcription-Factor Binding Patterns in t(8;21) Cells Correlate with the Response to RUNX1/ETO Knockdown

Genes bound by RUNX1/ETO are preferentially upregulated, whereas genes not bound by RUNX1/ETO are preferentially downregulated.

(A) Analysis of combinatorial binding identifies prevalent patterns in Kasumi-1 cells. The numbers of peaks are shown on the left of the heatmap for 61 factor-binding combinations (red: bound, scored as 1; blue: not bound, scored as 0 with the order of factors as depicted on top of the heatmap). *Z* scores on the right indicate the significance of deviation between observed and expected instances for all 61 combinatorial binding patterns. We identified 11 overrepresented binding patterns, which we analyzed further when each was associated with more than 100 genes. GSEA of selected large groups of genes (indicated by arrows) shows a highly significant enrichment of genes upregulated (upper left) or downregulated (lower left) after 4 days of RUNX1/ETO knockdown.

(B) Heatmap showing the RNA-seq overall fold change in Kasumi-1 cells 4 days after RUNX1/ETO knockdown.

(C) GSEA plots showing enrichment for up- or downregulated genes associated with dominant binding patterns in patient cells subjected to RUNX1/ETO knockdown, demonstrating that changes in gene expression were concordant between Kasumi-1 and patient cells after RUNX1/ETO knockdown. Note that in patient cells, RUNX1/ETO was only depleted for 48 hr and it takes about 4 days for the majority of genes to be downregulated (Ptasinska et al., 2012), thus explaining the lower p value seen with downregulated genes.

See also Figure S4.

Using the different overrepresented binding patterns, we constructed an interacting transcriptional network (Figure S4D). Most genes were regulated by a single binding pattern (node), and only some of these genes were associated with *cis* elements that bound different factor combinations (depicted as located between nodes). This specific binding pattern is of biological relevance because the genes that occupied the different network nodes clustered by overlapping but distinct Gene Ontology (GO) terms and KEGG pathways (Figures S4D and S4F; Table S2), indicating that they perform different functions. For example, *cis*-regulatory elements that bind RUNX1/ETO and all other factors (pattern 111111) are associated with genes involved in myeloid differentiation and hematopoiesis (Figure S4E; Table S2). Among the genes without RUNX1/ETO binding (pattern 011111) that were downregulated after RUNX1/ETO knockdown, we found the transcription factor genes *ERG* and *ETV6* (*TEL1*) (Figure S4F; Table S2), both of which are important for stem cell function and maintenance (Taoudi et al., 2011; Wang et al., 1998) but also have been implicated in AML (Diffner et al., 2013). ERG has also been shown to be important for stabilization of the RUNX1/ETO complex (Martens et al., 2012). Another downregulated transcription factor gene was *MEF2C*, which encodes a transcription factor that modulates myeloid fate and has oncogenic activity when overexpressed (Schwieger et al., 2009).

In summary, our analysis of the RUNX1/ETO-responsive core transcriptional network in t(8;21) cells highlights the predominantly repressive role of RUNX1/ETO within this network. Moreover, our analysis identified distinct classes of genes, with repressed genes involved in myeloid differentiation and active genes forming part of the stem cell signature.

Knockdown of RUNX1/ETO Leads to a Dynamic Reorganization of Transcription-Factor Binding

We next examined how the t(8;21) core transcriptional network changed 2 days after RUNX1/ETO depletion. Depletion had no immediate influence on the expression levels of any of the other factors studied above, with the notable exception of C/EBPα (Figure 3A). Nevertheless, loss of RUNX1/ETO had a profound effect on the binding of these transcription factors (Figure S5A). As exemplified by the *CEBPE* locus, depletion led to increased RUNX1 occupancy at several thousand sites, confirming that RUNX1/ETO and RUNX1 binding are in equilibrium (Figures 5A, top left, 5B, S5B, and S5C). Furthermore, increased RUNX1 occupancy, including RUNX1 sites that were not previously bound by RUNX1/ETO, was associated with a strong increase in p300 binding (Figure 3D). In contrast, more than 3,000 LMO2 binding sites were lost, mainly outside the regions bound by RUNX1/ETO and RUNX1 (Figures 5A, bottom-right panel, and S5C). Furthermore, whereas 80% of all PU.1 binding sites remained unchanged, the number of sites bound by C/EBPα increased 4-fold. Interestingly, 65% of all C/EBPα de novo sites colocalized with PU.1 (Figures 5A, top left, S5B, and S5D). In agreement with these results, C/EBPα binding sites clustered more strongly with both RUNX1 and PU.1 sites upon depletion of RUNX1/ETO (Figure S5E).

The changes in RUNX1 and C/EBPα binding, however, were not reflected by major global changes in DHS patterns. The comparison of DHS profiles before and after 2 days of RUNX1/ETO knockdown revealed that the majority of DHSs were unchanged (Figure 5C). Both C/EBPα and RUNX1 mainly associated with DHSs that were already present before RUNX1/ETO depletion. Only 20% of sites showed increased DNaseI sensitivity or arose de novo following RUNX1/ETO knockdown coinciding with de novo RUNX1 and C/EBPα binding (Figures S5F and S5G).

In summary, knockdown of RUNX1/ETO led to immediate genome-wide alterations in transcription-factor binding after 48 hr. Although a small fraction of binding sites arose de novo, this reprogramming occurred predominantly within preexisting transcription-factor assemblies.

The Dynamic Reorganization of the Leukemic Transcriptional Network after RUNX1/ETO Depletion Is Driven by C/EBPα

Many transcription factors upregulate the expression of their own gene, with *PU.1* (*SPI1*) being a prominent example (Leddin et al., 2011; Staber et al., 2013). However, of all the transcription factors examined, only C/EBPα was found to be significantly increased after RUNX1/ETO depletion (Figure 3A). Similarly to PU.1, C/EBPα upregulates its own expression in murine cells, and it was previously suggested that RUNX1/ETO interferes with C/EBPα expression by sequestering it from its promoter and thereby suppressing autoactivation (Pabst et al., 2001a). Our data demonstrate binding of C/EBPα to an element about 40 kb downstream of its own gene, a site that is also occupied by RUNX1/ETO, suggesting a more direct mechanism of repression (Ptasinska et al., 2012). C/EBPα is absolutely essential for terminal myeloid differentiation (Zhang et al., 1997) and occupies a large number of binding sites in mature macrophages (Heinz et al., 2010). However, *CEBPA* is not the only direct target gene of the *CEBP* family that responds to RUNX1/ETO: *CEBPE* and *CEBPD* are upregulated as well (Ptasinska et al., 2012), indicating that these factors may be part of a wider network of C/EBP proteins that control myeloid gene expression.

To test whether increased expression of C/EBPα was crucially involved in shifting the transcriptional network after RUNX1/ETO depletion, we defined overrepresented binding patterns for C/EBPα, PU.1, RUNX1, and LMO2 after RUNX1/ETO knockdown. Loss of RUNX1/ETO resulted in the formation of a transcriptional network dominated by C/EBPα-containing binding patterns, all of which were predominantly associated with upregulated genes in RUNX1/ETO-depleted Kasumi-1 and patient cells (Figures 6A–6C, S6A, and S6B; Table S3). Different patterns were again indicative of different classes of genes in terms of both GO and pathway analyses, with differentiation and signal transduction pathways being prominently featured (Figures S6C, S6D, and S7A). However, increased C/EBPα binding was also observed with a subset of genes that were downregulated (Figure 6D). Previous studies have shown that in addition to C/EBPα's role in driving myeloid differentiation, low levels of C/EBPα are required for stem cell maintenance, as upregulation of C/EBPα represses genes required for stem-cell self-renewal (Zhang et al., 2004, 2013). Therefore, we identified genes that (1) were downregulated after RUNX1/ETO knockdown and (2) showed increased C/EBPα binding (a total of 145 genes met

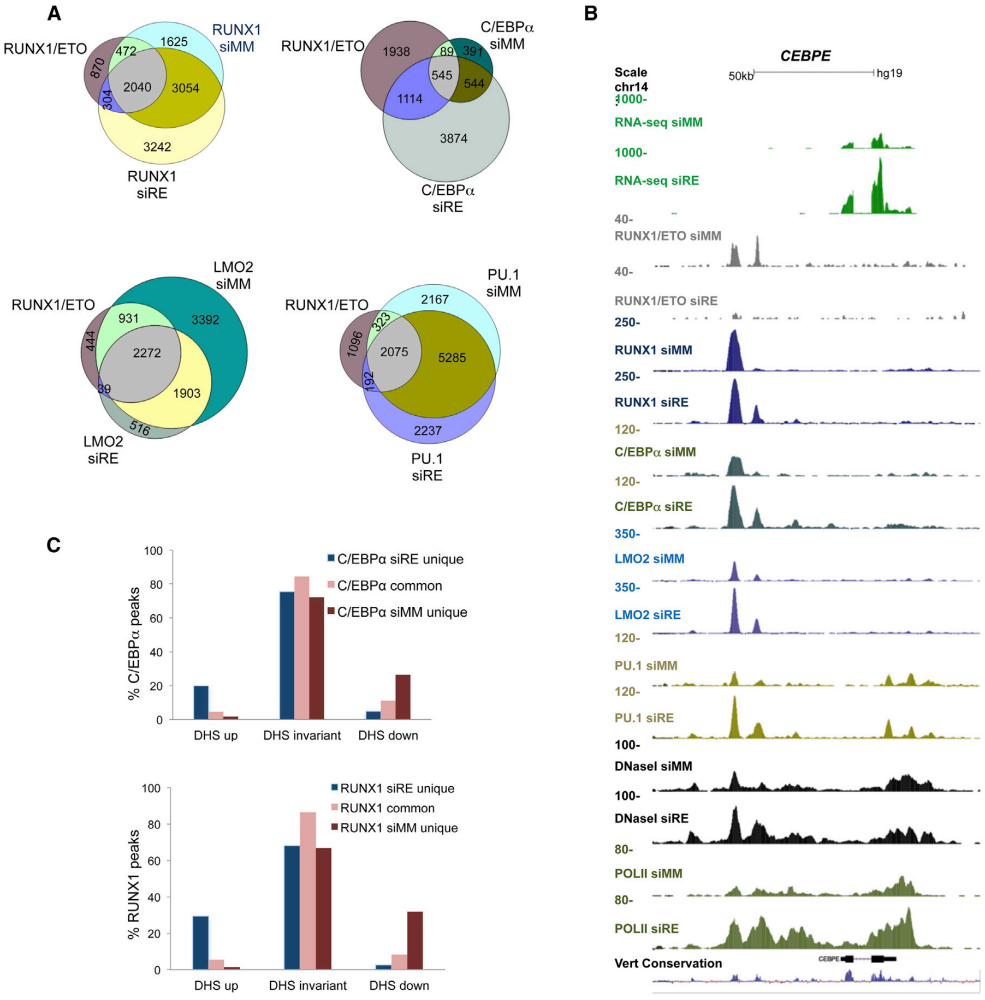


Figure 5. Knockdown of RUNX1/ETO Leads to a Reorganization of Transcription-Factor Assemblies within Preexisting Open Chromatin Regions

(A) Three-way Venn diagrams showing the overlap between RUNX1/ETO and RUNX1 (top left), CEBPα (top right), LMO2 (bottom left), and PU.1 (bottom right) in Kasumi-1 cells treated for 48 hr with control (siMM) and with RUNX1/ETO siRNA (siRE).

(B) UCSC genome browser screenshot showing the binding pattern of the indicated factors at the *CEBPE* locus in Kasumi-1 cells treated for 48 hr with control siRNA (siMM) and with RUNX1/ETO siRNA (siRE).

(C) Binding of de novo (siRE unique), common, and lost (siMM unique) transcription factors (C/EBPα (top) and RUNX1 (bottom) to regions of increased (DHS up), unchanged (DHS invariant), or reduced DNaseI hypersensitivity.

See also Figure S5.

the latter criterion; Figure 6D). This category included stem cell genes such as *ERG* and *CD34* (Figures S6F and S6G), as well as a large number of genes encoding for signaling molecules that are involved in regulating proliferation and differentiation, such as *DUSP6* or *PTK2* (Figure S6G).

We next evaluated whether C/EBPα was required for the upregulation of repressed RUNX1/ETO target genes. For this purpose, we depleted RUNX1/ETO with and without a concomitant C/EBPα knockdown. Knockdown of RUNX1/ETO led to a 2-fold increase in *C/EBPα* expression (Figures 3A, 7A, and 7B) and increases in expression of the direct RUNX1/ETO target genes, including *MS4A3*, *NKG7*, and *RNASE2*, which all show increased C/EBPα binding upon RUNX1/ETO depletion (Figures 7C and 7D; data not shown). Codepletion of C/EBPα diminished the induction of the three target genes in both Kasumi-1 and SKNO-1 cells (Figures 7D and S7B–S7D). These data indicate that

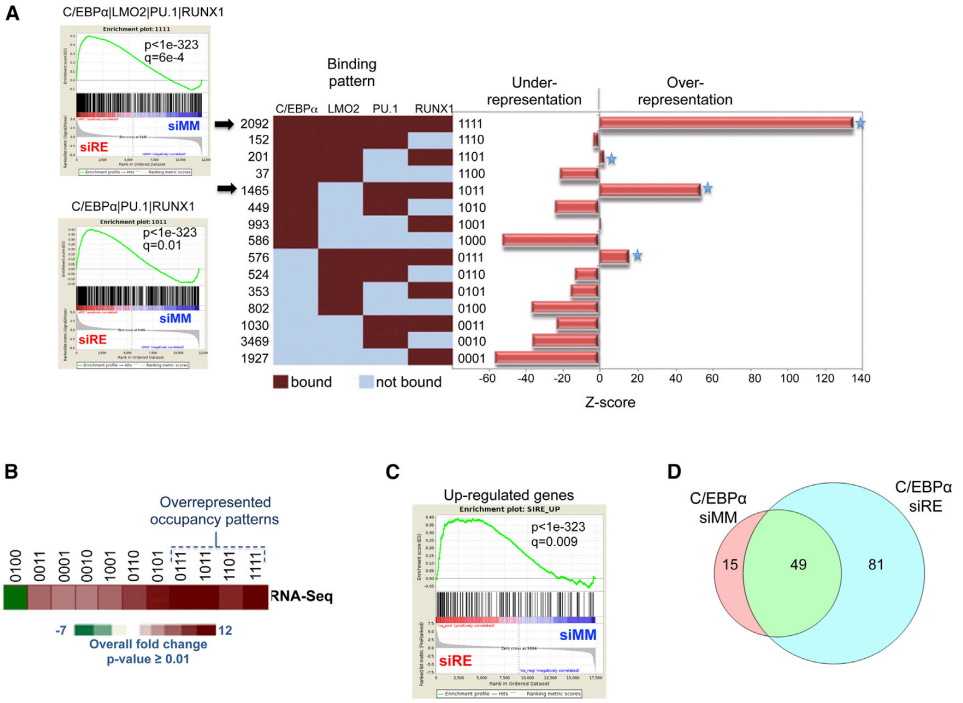


Figure 6. Transcriptional Network after RUNX1/ETO Depletion Is Enriched for C/EBPα Target Genes

(A) The transcription-factor binding state for CEBPα, LMO2, PU.1, and RUNX1 after RUNX1/ETO knockdown is characterized by an overrepresentation of four dominant occupancy patterns. The number of peaks for all 15 factor combinations is shown on the left of the heatmap (red: bound, scored as 1; blue: not bound, scored as 0). *Z* scores on the right indicate the significance of deviation between observed and expected instances for all 15 binding patterns. Left: GSEAs of genes associated with the two most enriched dominant occupancy patterns (indicated by arrows) show highly significant enrichment of upregulated genes after RUNX1/ETO knockdown.

(B) Genes associated with specific occupancy patterns that significantly change expression as measured by RNA-seq 4 days after RUNX1/ETO knockdown. The heatmap shows the RNA-seq overall fold change in Kasumi-1 cells 4 days after RUNX1/ETO knockdown.

(C) GSEAs showing that genes associated with dominant occupancy patterns that are upregulated in Kasumi-1 cells behave similarly in patient cells.

(D) Venn diagram depicting the number of genes bound by C/EBPα that are downregulated after RUNX1/ETO knockdown and show increased C/EBPα binding. See also Figure S6.

derepression of C/EBPα caused by RUNX1/ETO depletion is required for the full upregulation of a number of RUNX1/ETO target genes. However, we cannot rule out a similar function for other C/EBP members and in particular C/EBPδ and C/EBPε, which are both upregulated upon RUNX1/ETO knockdown (Figure 5B and data not shown). Nevertheless, our data confirm that C/EBPα plays an important role in orchestrating a transcriptional network that drives myeloid differentiation downstream of the original RUNX1/ETO network (Figure 7E).

DISCUSSION

The study presented here shows that expression of the oncogenic transcription factor RUNX1/ETO interferes with the hierarchical succession of transcriptional networks required for myeloid differentiation. Binding of RUNX1/ETO to key regulatory elements inhibits the expression of genes that drive differentiation. Moreover, we show that the establishment of a stable leukemic state not only depends on a static interaction of transcription factor complexes but also contains a dynamic competitive component as its key feature. We demonstrate that the transcriptional network controlled by RUNX1/ETO depends on a dynamic equilibrium between RUNX1/ETO and RUNX1 complexes, whose binding to their target sites is mutually exclusive. Although these complexes share the factors LMO2, HEB, and LYL1, they differ in their preferences for histone modifiers. RUNX1 can also act as a repressor (Levanon et al., 1998; Reed-Inderbitzin et al., 2006; Taniuchi et al., 2002), but in this factor context it preferentially recruits the HAT p300, whereas RUNX1/ETO recruits histone deacetylases, including HDAC2. RUNX1/ETO shares almost three-quarters of its binding sites with RUNX1, suggesting that the equilibrium between these two complexes results in a finely tuned modulation of expression for a wide range of genes. Thus, the leukemic phenotype requires the downmodulation of genes associated with differentiation, but may not tolerate their complete suppression.

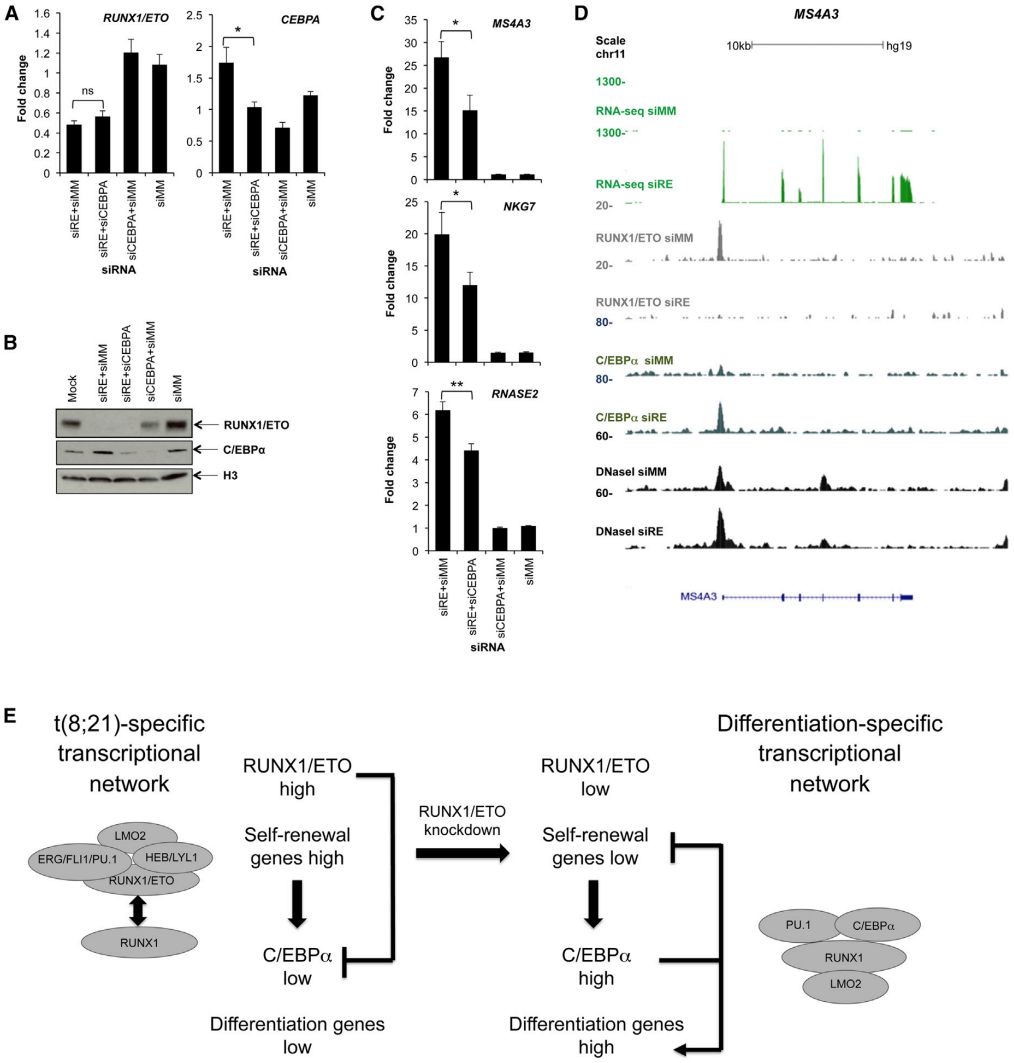


Figure 7. Loss of RUNX1/ETO Triggers C/EBPα-Driven Reorganization of the Leukemic Transcriptional Network

(A) *RUNX1/ETO* and *CEBPA* mRNA expression levels in Kasumi-1 cells 72 hr after electroporation with the indicated siRNAs. siRE, RUNX1/ETO siRNA; siCEBPA, C/EBPα siRNA; siMM, mismatch control siRNA. Results represent the mean ± SEM of five independent experiments. *$p < 0.05$; ns, not significant by paired Student's t test.

(B) Western blot indicating RUNX1/ETO and C/EBPα protein expression levels in single- and double-knockdown cells as indicated. An antibody against H3 was used as control. Mock, no siRNA.

(C) mRNA levels of *MS4A3*, *NKG7*, and *RNASE2* 72 hr after electroporation with the indicated siRNAs. Results represent the mean ± SEM of five independent experiments. *$p < 0.05$, **$p < 0.01$ by paired Student's t test.

(D) UCSC genome browser screenshot showing the binding pattern of RUNX1/ETO, C/EBPα, and DHSs at the *MS4A3* locus in Kasumi-1 cells treated for 48 hr with mismatch control siRNA (siMM) and with RUNX1/ETO siRNA (siRE).

(E) Model of RUNX1/ETO-mediated control of leukemic transcription. The competitive equilibrium in locus occupation between RUNX1/ETO and RUNX1 complexes drives leukemic self-renewal. Depletion of RUNX1/ETO increases the levels and DNA binding of its direct target gene, C/EBPα, which together with other differentiation genes reinstalls a transcriptional program that promotes myeloid differentiation.

See also Figure S7.

Consequently, perturbation of this equilibrium by depletion of RUNX1/ETO leads to loss of self-renewal, whereas knockdown of RUNX1 severely impairs viability (Ben-Ami et al., 2013; Dunne et al., 2006; Martinez et al., 2004; Martinez Soria et al., 2009). Currently, we do not know whether the different complexes exist independently or are in a rapid exchange. Evidence for both mechanisms exists; for example, in a previous study (Sun et al., 2013), neither p300 nor HDACs could be purified together with the RUNX1/ETO complex from t(8;21) cells using high stringency conditions. However, immunohistochemistry has demonstrated that RUNX1 and RUNX1/ETO are targeted to different subnuclear compartments (McNeil et al., 1999), a scenario that would be difficult to reconcile with a rapid exchange of factors binding to the same region of chromatin. Whatever the mechanism, it is likely that a mutually exclusive binding pattern can be found in other CBF leukemias. A similar colocalization with RUNX1 and its mutated counterpart has also been seen in AML with inversion 16 carrying the CBFB-MYH11 fusion protein (Mandoli et al., 2014), and furthermore, this type of AML is also dependent on the presence of an active copy of RUNX1 (Ben-Ami et al., 2013).

It was recently shown that aberrant RUNX1 expression is required for the maintenance of epithelial cancers (Scheitz et al., 2012). Moreover, RUNX1 plays a tumor-suppressive role by interacting with estrogen receptor α, and ERα-positive breast cancer patients carry mutations that disrupt these interactions (Chimge and Frenkel, 2013; Stender et al., 2010), highlighting increasing evidence that this factor and its deregulation or mutation are at the heart of multiple pathological processes. Moreover, alternative splicing of *RUNX1* leads to a C-terminally truncated isoform known as AML1a, which lacks the transactivation domain and promotes self-renewal of hematopoietic stem cells (Tsuzuki and Seto, 2012). We previously showed that during blood cell development, RUNX1 binding reshapes the epigenetic landscape by attracting other factors to its binding sites, and that this factor relocation is reversible (Lichtinger et al., 2012). Therefore, a dynamic equilibrium between different RUNX1 isoforms and other factors may also be relevant for cancers outside of the hematopoietic system.

A second important finding of our study is that the destruction of the RUNX1/ETO network establishes a transcription network dominated by the combinatorial binding of PU.1, RUNX1, and, in particular, C/EBPα (Figure 7E). Once RUNX1/ETO is depleted, C/EBPα expression levels increase and this factor then occupies a large number of binding sites, demonstrating at the genome-wide level that (1) C/EBPα is a major driver of myeloid differentiation and (2) the differentiation block in AML is partly caused by C/EBPα downregulation. The latter observation is consistent with the fact that a large number of AMLs involve mutations of C/EBPα (Preudhomme et al., 2002). However, the majority of binding sites are found in regions of previously accessible chromatin, indicating that (1) RUNX1/ETO targets binding sites that are destined for differentiation-driven factor exchange, and (2) shortly after its upregulation, C/EBPα resumes its original binding behavior and reorganizes existing transcription factor assemblies to drive myelopoiesis. These results tie in with the finding that PU.1 binding was largely invariant before and after RUNX1/ETO depletion. Although previous overexpression experiments indicated that RUNX1/ETO inactivated PU.1 (Vangala et al., 2003), our data indicate that, at least during the time window of 2 days, the PU.1 cistrome is largely unperturbed by the presence or absence of RUNX1/ETO and forms a platform upon which other factors dynamically assemble (Natoli et al., 2011).

In summary, our work sheds light on global mechanisms of the differentiation block in t(8;21) AML, which is of conceptual relevance for other types of AML and even other cancers. Many AML types are characterized by mutations in C/EBPα and RUNX1, which would impact many of the binding sites described here. The dynamic equilibrium between a mutated transcription factor and its wild-type counterpart allows a rapid reversion from a transcriptional program promoting malignant self-renewal to a differentiation program. Such dynamic behavior is likely to be the molecular cause of the good prognosis of t(8;21) AML and may also be a major angle for therapeutic intervention in other types of AML without mutations in other hematopoietic regulators.

EXPERIMENTAL PROCEDURES

More detailed descriptions of the materials and methods used in this work can be found in the Supplemental Experimental Procedures.

Human Patient Cells and Cell Lines

Patient material was obtained with approval from the NHS Research Ethics Committees (Leeds Teaching Hospitals NHS Trust and Newcastle upon Tyne Hospitals NHS Foundation Trust). Kasumi-1 cells were obtained from the DSMZ cell line repository (http://www.dsmz.de/) and were cultured in RPMI1640 containing 10% fetal calf serum (FCS). SKNO-1 cells were maintained in RPMI1640 supplemented with 20% FCS and 7 ng/ml granulocyte-macrophage colony-stimulating factor.

siRNA Transfections

Kasumi-1 and SKNO-1 cells were transfected with 200 nM siRNA using a Fischer EPI 3500 electroporator (Fischer) as described previously (Ptasinska et al., 2012). The following siRNAs were used: RUNX1/ETO siRNA (sense, CCUCGAAAUCGUACUGAGAAG; antisense, UCUCAGUACGAUUUCGAGGUU), mismatch control siRNA (sense, CCUCGAAUUCGUUCUGAGAAG; antisense, UC UCAGAACGAAUUCGAGGUU); and C/EBPα siRNA (sense, CCG GAGUUAUGACAAGCUUUC; antisense, AAGCUUGUCAUAACUCCGGUC).

Real-Time RT-PCR

RNA extraction and quantitative real-time RT-PCR were performed as described previously (Ptasinska et al., 2012). Primers are listed in Table S4.

Western Blotting

Kasumi-1 cells were lysed in RIPA buffer 2 days after electroporation. The following antibodies were used for western blot analysis: C/EBPα, ab15048 (Abcam); ETO, SC-9737 (Santa Cruz Biotechnology); GAPDH, ab8245 (Abcam); HDAC2, ab7029 (Abcam); HEB, SC-357 (Santa Cruz); LDB1, SC-11198 (Santa Cruz); LMO2, AF2726 (R&D Systems); LYL1, SC-374164 (Santa Cruz); PU.1, SC-352 (Santa Cruz); p300, SC-585 (Santa Cruz); and RUNX1, PC285 (Millipore).

ChIP

ChIP assays were performed as described previously (Ptasinska et al., 2012). Nuclei were essentially prepared as described previously (Lefevre et al., 2003). The following antibodies were used: C/EBPα, SC-61 (Santa Cruz Biotechnology); ETO (C terminus specific), SC-9737 (Santa Cruz); HDAC2, SC-6296 (Santa Cruz); HEB, SC-357 (Santa Cruz); LMO2, AF2726 (R&D Systems); LYL1, SC-374164 (Santa Cruz); PU.1, SC-352 (Santa Cruz); p300, SC-585 (Santa Cruz); RUNX1 (C terminus specific), ab23980 (Abcam) or IgG rabbit 12-370 (Millipore); IgG goat, SC-2346 (Santa Cruz); and IgG mouse,

SC-2025 (Santa Cruz). Precipitated material was subjected to library preparation and run on an Illumina Hiseq 2000 sequencer.

RNA-Seq

RNA samples from three independent biological replicates were processed using the Tru-seq RNA Sample Prep Kit v2 (Illumina) according to the manufacturer's protocol. Libraries were run in 4× multiplex on an Illumina Hiseq 2000 sequencer generating ~90 million paired-end reads per sample.

Re-ChIP

Re-ChIP was carried out as described above with minor modifications. Following the final ChIP wash, chromatin complexes were eluted twice in 50 μl of ChIP elution buffer (100 mM $NaHCO_3$, 1% SDS, PIC) for 15 min at room temperature with shaking. Eluates were combined and diluted 20 times with ChIP dilution buffer, followed by a 5 hr incubation with the second primary antibody or IgG. After elution with 100 mM $NaHCO_3$, 1% SDS for 30 min at room temperature, the re-ChIP products were analyzed by quantitative PCR (qPCR). Fold-enrichment values were calculated relative to a negative control region of the genome. Primers are listed in Table S4.

DHS Mapping

Genome-wide DHSs were mapped as described previously (Leddin et al., 2011).

Library Generation and Sequencing

Libraries of DNA fragments from ChIP or DNase I treatment were prepared from 10 ng of DNA according to standard procedures. ETO, RUNX1, C/EBPα, PU.1, LMO2 ChIP, and Kasumi-1 DNase I libraries were sequenced on an Illumina Genome Analyzer GAIIx using 36 bp single-end reads. For patients 1 and 2, DNase I (491 and 342 million reads, respectively) and control patient libraries (Table S1) were sequenced on an Illumina HiSeq using 50 bp single-end reads.

ACCESSION NUMBERS

The GEO accession numbers for the data reported in this paper are GSE29225 (Ptasinska et al., 2012) and GSE54478.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.08.024.

AUTHOR CONTRIBUTIONS

A.P., M.R.I., N.M.-S., A.P., M.W., S.J., and M.H. performed experiments. S.A.A., P.C., J.P., S.O., D.R.W., and D.W. analyzed data. D.G.T. provided technical infrastructure and helped write the manuscript. P.N.C. supervised experiments and helped write the manuscript. C.B. and O.H. conceived the study, supervised experiments, and wrote the manuscript.

ACKNOWLEDGMENTS

The authors thank Simon Bomken, Luke Gaughan, and John Lunec for carefully reading and improving the manuscript. Research in the C.B. lab is supported by grants from Leukaemia & Lymphoma Research (7001 and 12007) and the Medical Research Council, UK. O.H. received support from Leukaemia & Lymphoma Research (10033 and 12055) and the North of England Children's Cancer Fund.

Received: January 31, 2014
Revised: June 19, 2014
Accepted: August 12, 2014
Published: September 18, 2014

REFERENCES

Amann, J.M., Nip, J., Strom, D.K., Lutterbach, B., Harada, H., Lenny, N., Downing, J.R., Meyers, S., and Hiebert, S.W. (2001). ETO, a target of t(8;21) in acute leukemia, makes distinct contacts with multiple histone deacetylases and binds mSin3A through its oligomerization domain. Mol. Cell. Biol. *21*, 6470–6483.

Ben-Ami, O., Friedman, D., Leshkowitz, D., Goldenberg, D., Orlovsky, K., Pencovich, N., Lotem, J., Tanay, A., and Groner, Y. (2013). Addiction of t(8;21) and inv(16) acute myeloid leukemia to native RUNX1. Cell Rep *4*, 1131–1143.

Chimge, N.O., and Frenkel, B. (2013). The RUNX family in breast cancer: relationships with estrogen signaling. Oncogene *32*, 2121–2130.

Davidson, E.H. (2010). Emerging properties of animal gene regulatory networks. Nature *468*, 911–920.

DeVilbiss, A.W., Sanalkumar, R., Johnson, K.D., Keles, S., and Bresnick, E.H. (2014). Hematopoietic transcriptional mechanisms: From locus-specific to genome-wide vantage points. Exp. Hematol. *42*, 618–629.

Diffner, E., Beck, D., Gudgin, E., Thoms, J.A., Knezevic, K., Pridans, C., Foster, S., Goode, D., Lim, W.K., Boelen, L., et al. (2013). Activity of a heptad of transcription factors is associated with stem cell programs and clinical outcome in acute myeloid leukemia. Blood *121*, 2289–2300.

Dunne, J., Cullmann, C., Ritter, M., Soria, N.M., Drescher, B., Debernardi, S., Skoulakis, S., Hartmann, O., Krause, M., Krauter, J., et al. (2006). siRNA-mediated AML1/MTG8 depletion affects differentiation and proliferation-associated gene expression in t(8;21)-positive cell lines and primary AML blasts. Oncogene *25*, 6067–6078.

Follows, G.A., Tagoh, H., Lefevre, P., Hodge, D., Morgan, G.J., and Bonifer, C. (2003). Epigenetic consequences of AML1-ETO action at the human c-FMS locus. EMBO J. *22*, 2798–2809.

Gaidzik, V.I., Bullinger, L., Schlenk, R.F., Zimmermann, A.S., Röck, J., Paschka, P., Corbacioglu, A., Krauter, J., Schlegelberger, B., Ganser, A., et al. (2011). RUNX1 mutations in acute myeloid leukemia: results from a comprehensive genetic and clinical analysis from the AML study group. J. Clin. Oncol. *29*, 1364–1372.

Goyama, S., Schibler, J., Cunningham, L., Zhang, Y., Rao, Y., Nishimoto, N., Nakagawa, M., Olsson, A., Wunderlich, M., Link, K.A., et al. (2013). Transcription factor RUNX1 promotes survival of acute myeloid leukemia cells. J. Clin. Invest. *123*, 3876–3888.

Heidenreich, O., Krauter, J., Riehle, H., Hadwiger, P., John, M., Heil, G., Vornlocher, H.P., and Nordheim, A. (2003). AML1/MTG8 oncogene suppression by small interfering RNAs supports myeloid differentiation of t(8;21)-positive leukemic cells. Blood *101*, 3157–3163.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors primecis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Kitabayashi, I., Yokoyama, A., Shimizu, K., and Ohki, M. (1998). Interaction and functional cooperation of the leukemia-associated factors AML1 and p300 in myeloid cell differentiation. EMBO J. *17*, 2994–3004.

Leddin, M., Perrod, C., Hoogenkamp, M., Ghani, S., Assi, S., Heinz, S., Wilson, N.K., Follows, G., Schönheit, J., Vockentanz, L., et al. (2011). Two distinct auto-regulatory loops operate at the PU.1 locus in B cells and myeloid cells. Blood *117*, 2827–2838.

Lefevre, P., Melnik, S., Wilson, N., Riggs, A.D., and Bonifer, C. (2003). Developmentally regulated recruitment of transcription factors and chromatin modification activities to chicken lysozymecis-regulatory elements in vivo. Mol. Cell. Biol. *23*, 4386–4400.

Levanon, D., Goldstein, R.E., Bernstein, Y., Tang, H., Goldenberg, D., Stifani, S., Paroush, Z., and Groner, Y. (1998). Transcriptional repression by AML1 and LEF-1 is mediated by the TLE/Groucho corepressors. Proc. Natl. Acad. Sci. USA *95*, 11590–11595.

Lichtinger, M., Ingram, R., Hannah, R., Müller, D., Clarke, D., Assi, S.A., Lie-A-Ling, M., Noailles, L., Vijayabaskar, M.S., Wu, M., et al. (2012). RUNX1

reshapes the epigenetic landscape at the onset of haematopoiesis. EMBO J. *31*, 4318–4333.

Liu, Y., Cheney, M.D., Gaudet, J.J., Chruszcz, M., Lukasik, S.M., Sugiyama, D., Lary, J., Cole, J., Dauter, Z., Minor, W., et al. (2006). The tetramer structure of the Nervy homology two domain, NHR2, is critical for AML1/ETO's activity. Cancer Cell *9*, 249–260.

Mandoli, A., Singh, A.A., Jansen, P.W., Wierenga, A.T., Riahi, H., Franci, G., Prange, K., Saeed, S., Vellenga, E., Vermeulen, M., et al. (2014). CBFB-MYH11/RUNX1 together with a compendium of hematopoietic regulators, chromatin modifiers and basal transcription factors occupies self-renewal genes in inv(16) acute myeloid leukemia. Leukemia *28*, 770–778.

Martens, J.H., Mandoli, A., Simmer, F., Wierenga, B.J., Saeed, S., Singh, A.A., Altucci, L., Vellenga, E., and Stunnenberg, H.G. (2012). ERG and FLI1 binding sites demarcate targets for aberrant epigenetic regulation by AML1-ETO in acute myeloid leukemia. Blood *120*, 4038–4048.

Martinez, N., Drescher, B., Riehle, H., Cullmann, C., Vornlocher, H.P., Ganser, A., Heil, G., Nordheim, A., Krauter, J., and Heidenreich, O. (2004). The oncogenic fusion protein RUNX1-CBFA2T1 supports proliferation and inhibits senescence in t(8;21)-positive leukaemic cells. BMC Cancer *4*, 44.

Martinez Soria, N., Tussiwand, R., Ziegler, P., Manz, M.G., and Heidenreich, O. (2009). Transient depletion of RUNX1/RUNX1T1 by RNA interference delays tumour formation in vivo. Leukemia *23*, 188–190.

McNeil, S., Zeng, C., Harrington, K.S., Hiebert, S., Lian, J.B., Stein, J.L., van Wijnen, A.J., and Stein, G.S. (1999). The t(8;21) chromosomal translocation in acute myelogenous leukemia modifies intranuclear targeting of the AML1/CBFalpha2 transcription factor. Proc. Natl. Acad. Sci. USA *96*, 14882–14887.

Michaud, J., Wu, F., Osato, M., Cottles, G.M., Yanagida, M., Asou, N., Shigesada, K., Ito, Y., Benson, K.F., Raskind, W.H., et al. (2002). In vitro analyses of known and novel RUNX1/AML1 mutations in dominant familial platelet disorder with predisposition to acute myelogenous leukemia: implications for mechanisms of pathogenesis. Blood *99*, 1364–1372.

Miyoshi, H., Kozu, T., Shimizu, K., Enomoto, K., Maseki, N., Kaneko, Y., Kamada, N., and Ohki, M. (1993). The t(8;21) translocation in acute myeloid leukemia results in production of an AML1-MTG8 fusion transcript. EMBO J. *12*, 2715–2721.

Natoli, G., Ghisletti, S., and Barozzi, I. (2011). The genomic landscapes of inflammation. Genes Dev. *25*, 101–106.

Pabst, T., Mueller, B.U., Harakawa, N., Schoch, C., Haferlach, T., Behre, G., Hiddemann, W., Zhang, D.E., and Tenen, D.G. (2001a). AML1-ETO downregulates the granulocytic differentiation factor C/EBPalpha in t(8;21) myeloid leukemia. Nat. Med. *7*, 444–451.

Pabst, T., Mueller, B.U., Zhang, P., Radomska, H.S., Narravula, S., Schnittger, S., Behre, G., Hiddemann, W., and Tenen, D.G. (2001b). Dominant-negative mutations of CEBPA, encoding CCAAT/enhancer binding protein-alpha (C/EBPalpha), in acute myeloid leukemia. Nat. Genet. *27*, 263–270.

Pimanda, J.E., and Göttgens, B. (2010). Gene regulatory networks governing haematopoietic stem cell development and identity. Int. J. Dev. Biol. *54*, 1201–1211.

Piper, J., Elze, M.C., Cauchy, P., Cockerill, P.N., Bonifer, C., and Ott, S. (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Res. *41*, e201.

Preudhomme, C., Sagot, C., Boissel, N., Cayuela, J.M., Tigaud, I., de Botton, S., Thomas, X., Raffoux, E., Lamandin, C., Castaigne, S., et al.; ALFA Group (2002). Favorable prognostic significance of CEBPA mutations in patients with de novo acute myeloid leukemia: a study from the Acute Leukemia French Association (ALFA). Blood *100*, 2717–2723.

Ptasinska, A., Assi, S.A., Mannari, D., James, S.R., Williamson, D., Dunne, J., Hoogenkamp, M., Wu, M., Care, M., McNeill, H., et al. (2012). Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding. Leukemia *26*, 1829–1841.

Reed-Inderbitzin, E., Moreno-Miralles, I., Vanden-Eynden, S.K., Xie, J., Lutterbach, B., Durst-Goodwin, K.L., Luce, K.S., Irvin, B.J., Cleary, M.L., Brandt, S.J., and Hiebert, S.W. (2006). RUNX1 associates with histone deacetylases and SUV39H1 to repress transcription. Oncogene *25*, 5777–5786.

Saeed, S., Logie, C., Francoijs, K.J., Frigè, G., Romanenghi, M., Nielsen, F.G., Raats, L., Shahhoseini, M., Huynen, M., Altucci, L., et al. (2012). Chromatin accessibility, p300, and histone acetylation define PML-RARα and AML1-ETO binding sites in acute myeloid leukemia. Blood *120*, 3058–3068.

Scheitz, C.J., and Tumbar, T. (2013). New insights into the role of Runx1 in epithelial stem cell biology and pathology. J. Cell. Biochem. *114*, 985–993.

Scheitz, C.J., Lee, T.S., McDermitt, D.J., and Tumbar, T. (2012). Defining a tissue stem cell-driven Runx1/Stat3 signalling axis in epithelial cancer. EMBO J. *31*, 4124–4139.

Schwieger, M., Schüler, A., Forster, M., Engelmann, A., Arnold, M.A., Delwel, R., Valk, P.J., Löhler, J., Slany, R.K., Olson, E.N., and Stocking, C. (2009). Homing and invasiveness of MLL/ENL leukemic cells is regulated by MEF2C. Blood *114*, 2476–2488.

Scott, E.W., Simon, M.C., Anastasi, J., and Singh, H. (1994). Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. Science *265*, 1573–1577.

Snaddon, J., Smith, M.L., Neat, M., Cambal-Parrales, M., Dixon-McIver, A., Arch, R., Amess, J.A., Rohatiner, A.Z., Lister, T.A., and Fitzgibbon, J. (2003). Mutations of CEBPA in acute myeloid leukemia FAB types M1 and M2. Genes Chromosomes Cancer *37*, 72–78.

Staber, P.B., Zhang, P., Ye, M., Welner, R.S., Nombela-Arrieta, C., Bach, C., Kerenyi, M., Bartholdy, B.A., Zhang, H., Alberich-Jordà, M., et al. (2013). Sustained PU.1 levels balance cell-cycle regulators to prevent exhaustion of adult hematopoietic stem cells. Mol. Cell *49*, 934–946.

Stender, J.D., Kim, K., Charn, T.H., Komm, B., Chang, K.C., Kraus, W.L., Benner, C., Glass, C.K., and Katzenellenbogen, B.S. (2010). Genome-wide analysis of estrogen receptor alpha DNA binding and tethering mechanisms identifies Runx1 as a novel tethering factor in receptor-mediated transcriptional activation. Mol. Cell. Biol. *30*, 3943–3955.

Sun, X.J., Wang, Z., Wang, L., Jiang, Y., Kost, N., Soong, T.D., Chen, W.Y., Tang, Z., Nakadai, T., Elemento, O., et al. (2013). A stable transcription factor complex nucleated by oligomeric AML1-ETO controls leukaemogenesis. Nature *500*, 93–97.

Taniuchi, I., Osato, M., Egawa, T., Sunshine, M.J., Bae, S.C., Komori, T., Ito, Y., and Littman, D.R. (2002). Differential requirements for Runx proteins in CD4 repression and epigenetic silencing during T lymphocyte development. Cell *111*, 621–633.

Taoudi, S., Bee, T., Hilton, A., Knezevic, K., Scott, J., Willson, T.A., Collin, C., Thomas, T., Voss, A.K., Kile, B.T., et al. (2011). ERG dependence distinguishes developmental control of hematopoietic stem cell maintenance from hematopoietic specification. Genes Dev. *25*, 251–262.

Tijssen, M.R., Cvejic, A., Joshi, A., Hannah, R.L., Ferreira, R., Forrai, A., Bellissimo, D.C., Oram, S.H., Smethurst, P.A., Wilson, N.K., et al. (2011). Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. Dev. Cell *20*, 597–609.

Tsuzuki, S., and Seto, M. (2012). Expansion of functionally defined mouse hematopoietic stem and progenitor cells by a short isoform of RUNX1/AML1. Blood *119*, 727–735.

Valk, P.J., Verhaak, R.G., Beijen, M.A., Erpelinck, C.A., Barjesteh van Waalwijk van Doorn-Khosrovani, S., Boer, J.M., Beverloo, H.B., Moorhouse, M.J., van der Spek, P.J., Löwenberg, B., and Delwel, R. (2004). Prognostically useful gene-expression profiles in acute myeloid leukemia. N. Engl. J. Med. *350*, 1617–1628.

van Riel, B., Pakozdi, T., Brouwer, R., Monteiro, R., Tuladhar, K., Franke, V., Bryne, J.C., Jorna, R., Rijkers, E.J., van Ijcken, W., et al. (2012). A novel complex, RUNX1-MYEF2, represses hematopoietic genes in erythroid cells. Mol. Cell. Biol. *32*, 3814–3822.

Vangala, R.K., Heiss-Neumann, M.S., Rangatia, J.S., Singh, S.M., Schoch, C., Tenen, D.G., Hiddemann, W., and Behre, G. (2003). The myeloid master regulator transcription factor PU.1 is inactivated by AML1-ETO in t(8;21) myeloid leukemia. Blood *101*, 270–277.

Wang, J., Hoshino, T., Redner, R.L., Kajigaya, S., and Liu, J.M. (1998). ETO, fusion partner in t(8;21) acute myeloid leukemia, represses transcription by interaction with the human N-CoR/mSin3/HDAC1 complex. Proc. Natl. Acad. Sci. USA *95*, 10860–10865.

Wang, L., Gural, A., Sun, X.J., Zhao, X., Perna, F., Huang, G., Hatlen, M.A., Vu, L., Liu, F., Xu, H., et al. (2011). The leukemogenicity of AML1-ETO is dependent on site-specific lysine acetylation. Science *333*, 765–769.

Wilson, N.K., Foster, S.D., Wang, X., Knezevic, K., Schütte, J., Kaimakis, P., Chilarska, P.M., Kinston, S., Ouwehand, W.H., Dzierzak, E., et al. (2010). Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell *7*, 532–544.

Zhang, D.E., Zhang, P., Wang, N.D., Hetherington, C.J., Darlington, G.J., and Tenen, D.G. (1997). Absence of granulocyte colony-stimulating factor signaling and neutrophil development in CCAAT enhancer binding protein alpha-deficient mice. Proc. Natl. Acad. Sci. USA *94*, 569–574.

Zhang, P., Iwasaki-Arai, J., Iwasaki, H., Fenyus, M.L., Dayaram, T., Owens, B.M., Shigematsu, H., Levantini, E., Huettner, C.S., Lekstrom-Himes, J.A., et al. (2004). Enhancement of hematopoietic stem cell repopulating capacity and self-renewal in the absence of the transcription factor C/EBP alpha. Immunity *21*, 853–863.

Zhang, H., Alberich-Jorda, M., Amabile, G., Yang, H., Staber, P.B., Di Ruscio, A., Welner, R.S., Ebralidze, A., Zhang, J., Levantini, E., et al. (2013). Sox4 is a key oncogenic target in C/EBPα mutant acute myeloid leukemia. Cancer Cell *11*, 575–588.